Airport Flights and Parked Cars Dataset & Analysis

1. Research goals:

    - **Is there any relation between the number of private cars parked and the total number of flights at Incheon International Airport.**

    - **Identify specific destination-bound flights that significantly impact the number of private cars parked.**

2. Business questions:

    - **Is there any correlation between the number of flights and the number of private cars parked at the airport?**

    - **Which flight or destination (Europe, America, Asian, etc.) has the most significant impact on traffic in the parking area ?**

3. Data Collection:

    - **Airport flight information**

    - **Parking lot utilization**

Basic Characteristics

Below are some of the characteristics of the obtained dataset –

- The collected data amounted to 3,990 records and 16 columns.
- Missing values constitute 39.02% of the overall data
- Summary of different scales of data is as follows –

Scale of Data	Count of columns
Nominal	7
Ordinal	1
Ratio	0
Interval	8
Grand Total	**16**

Column Description and Scale

Column Names	Data type	Scale	Description	Missing Values
Date	object	Interval	Date of flights	0.00%
Departure Time	object	Interval	Time of Departure	0.00%
Destination	category	Nominal	The name of destination city	0.00%
Airline	category	Nominal	The name of air carrier	0.27%
Flight name	category	Nominal	Flight Code (e.g. DL7892)	0.00%
Terminal	category	Nominal	The number of Incheon Terminal (e.g. T1, T2)	0.00%
Check-in Counter	Category	Nominal	The number of Counter (e.g. H01-H08)	0.00%
Gate	category	Ordinal	The number of Gate (e.g. 38)	0.00%
Flight Status	category	Nominal	The status of flight (e.g. Departure, Cancel)	0.00%
Codeshare	category	Nominal	Seat-sharing Flight among Alliances (e.g. Master, Slave)	99.82%

Date2	object	Interval	Date for counting parked-in cars	99.82%
T1 Short	int8	Interval	The number of parked in cars in the Short-term parking lot of Terminal 1	99.82%
T1 Long	int8	Interval	The number of parked in cars in the Long-term parking lot of Terminal 1	99.82%
T2 Short	int8	Interval	The number of parked in cars in Short-term parking lot of Terminal 2	99.82%
T2 Long	int8	Interval	The number of parked in cars in the Long-term parking lot of Terminal 2	99.82%
Total	int8	Interval	Total number of parked in cars	99.82%

Initial Datasets

Our project started with two separate datasets. The first dataset contains the records of 3,990 flights departing from March 12, 2023 to March 18, 2023 including the following information for each flight:

- Time and date
- Destination
- Airline and flight number
- Gate and Terminal Information
- Codeshare Information

The second dataset contained parking information for each hour from March 12, 2023 to March 18, 2023. Information was split into incoming and outgoing cars for the following terminal parking lots:

- T1 Short-term
- T1 Long-term
- T2 Short-term
- T1 Long-term

There is no missing flight or parking data in either dataset. There were no duplicate flight entries either in the flight dataset.

These two datasets will need to be combined for the analysis. The destinations will also be combined into different regions. Codeshare information is likely unneeded and will be dropped.

Hourly Analysis

- ➢ The number of flights and cars entering hourly was analyzed:

Data was combined by hour. The dataset was reduced to flights per hour with destination region and the number of cars going into the lot. Below is an example of the dataset, with some regions removed from the image in order to better see the data.

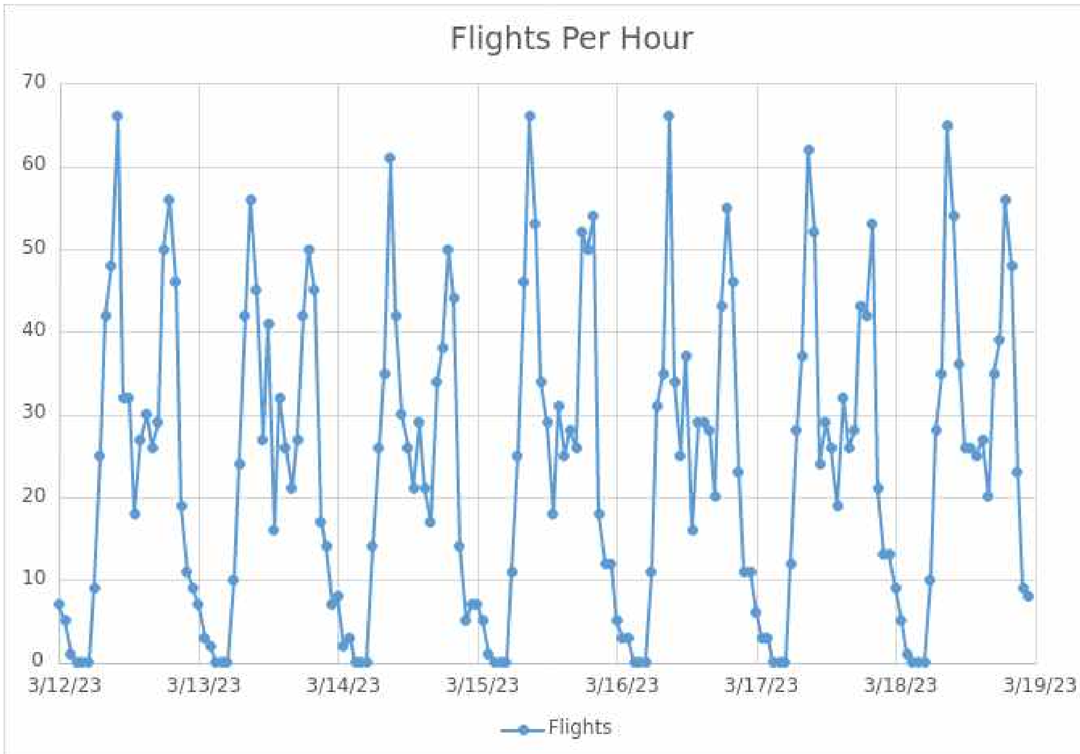
Time	Flights	T1 Short Term	T1 Long Term	T2 Short Term	T2 Long Term	Total Cars In	Europe	South Asia	Middle East	Africa	Vietnam	CentralAsia	Japan
3/12/2023 0:00	7	47	25	18	8	98	0	2	4	1	0	0	0
3/12/2023 1:00	5	24	24	6	4	58	1	0	0	0	3	1	0
3/12/2023 2:00	1	74	63	10	9	156	0	1	0	0	0	0	0
3/12/2023 3:00	0	116	143	26	36	321	0	0	0	0	0	0	0
3/12/2023 4:00	0	278	260	59	50	647	0	0	0	0	0	0	0
3/12/2023 5:00	0	565	546	327	123	1561	0	0	0	0	0	0	0
3/12/2023 6:00	9	757	487	567	166	1977	0	0	0	0	7	0	2
3/12/2023 7:00	25	718	466	479	95	1758	0	8	0	0	5	0	12
3/12/2023 8:00	42	628	393	406	56	1483	1	6	0	0	0	2	29
3/12/2023 9:00	48	506	185	226	44	961	1	4	0	0	4	0	24
3/12/2023 10:00	66	520	176	211	38	945	4	2	0	0	6	0	27

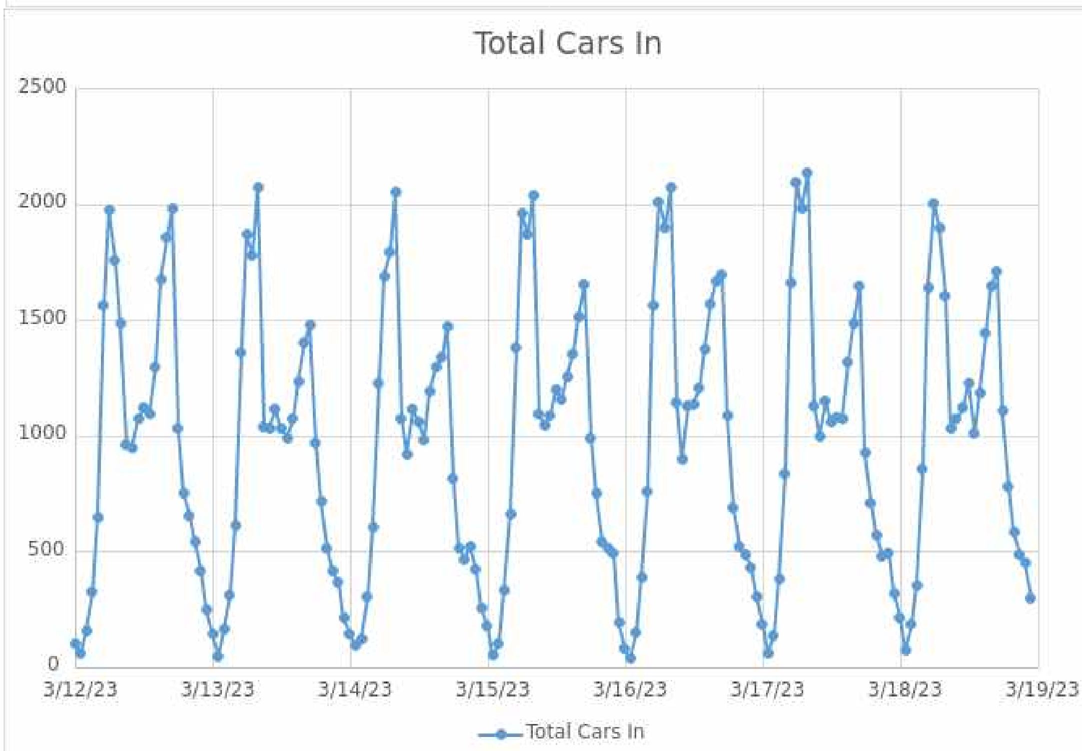
The descriptive statistics were generated for the flights and cars for each hour:

```
> summary(df)
     Time              Flights      T1.Short.Term    T1.Long.Term    T2.Short.Term
 Length:168         Min.   : 0.00   Min.   : 10.0   Min.   : 11.0   Min.   :  1.0
 Class :character   1st Qu.: 7.00   1st Qu.:278.0   1st Qu.:104.8   1st Qu.: 52.0
 Mode  :character   Median :25.00   Median :493.0   Median :235.5   Median :222.0
                    Mean   :23.75   Mean   :429.9   Mean   :258.6   Mean   :233.7
                    3rd Qu.:35.00   3rd Qu.:599.5   3rd Qu.:347.2   3rd Qu.:343.2
                    Max.   :66.00   Max.   :855.0   Max.   :728.0   Max.   :800.0
  T2.Long.Term    Total.Cars.In        Europe         South.Asia       Middle.East
 Min.   :  4.00   Min.   :  35.0   Min.   :0.0000   Min.   : 0.000   Min.   :0.0000
 1st Qu.: 11.75   1st Qu.: 475.2   1st Qu.:0.0000   1st Qu.: 1.000   1st Qu.:0.0000
 Median : 40.50   Median :1030.5   Median :0.0000   Median : 4.000   Median :0.0000
 Mean   : 44.97   Mean   : 967.2   Mean   :0.8988   Mean   : 4.958   Mean   :0.7083
 3rd Qu.: 62.00   3rd Qu.:1377.0   3rd Qu.:1.0000   3rd Qu.: 8.000   3rd Qu.:0.0000
 Max.   :166.00   Max.   :2135.0   Max.   :7.0000   Max.   :20.000   Max.   :8.0000
     Africa            Vietnam         CentralAsia         Japan           NearAsia
 Min.   :0.00000   Min.   : 0.000   Min.   :0.0000   Min.   : 0.000   Min.   : 0.000
 1st Qu.:0.00000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.: 0.000
 Median :0.00000   Median : 1.000   Median :0.0000   Median : 3.500   Median : 0.000
 Mean   :0.02381   Mean   : 2.815   Mean   :0.4881   Mean   : 6.899   Mean   : 1.958
 3rd Qu.:0.00000   3rd Qu.: 5.000   3rd Qu.:0.0000   3rd Qu.:11.250   3rd Qu.: 4.000
 Max.   :1.00000   Max.   :19.000   Max.   :5.0000   Max.   :33.000   Max.   :13.000
     China          PacificOcean       America            Aus              Korea
 Min.   :0.0000   Min.   :0.0000   Min.   : 0.000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median : 0.000   Median :0.0000   Median :0.0000
 Mean   :0.8512   Mean   :0.7262   Mean   : 2.673   Mean   :0.4167   Mean   :0.3333
 3rd Qu.:2.0000   3rd Qu.:0.0000   3rd Qu.: 5.000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :6.0000   Max.   :8.0000   Max.   :17.000   Max.   :6.0000   Max.   :4.0000
```

The typical hour has an average of 23.75 flights departing. An average of 967.2 cars enter the lots every hour. The T1 short term lot has the most cars entering per hour, at a rate of 429.9/hour. T2 Long term has the lowest rate at 44.97/hour. South Asia is the most frequent destination, at 5.0 flights per hour. Africa is the least frequent destination, at only 0.02 flights per hour.

- The hourly charts for flights and cars entering the lot per hour:





There appears to be a strong correlation between flights and cars entering the lots, as the two graphs peak at the same times each day.

Model Creation:

Step 1:

Terminal information was added to the combined dataset. This will allow the terminal of the flight to be used as a variable in the model. The number of cars are recorded for the separate lots and the flight terminal can be used in a model to determine if there is a correlation between the terminal of the flight, and where passengers park their cars.

Time	Flights	Flights T1	Flights T2	T1 Short Term	T1 Long Term	T2 Short Term	T2 Long Term	Total Cars In	Europe	South Asia	Middle East
3/12/2023 0:00	7	7	0	47	25	18	8	98	0	2	4
3/12/2023 1:00	5	4	1	24	24	6	4	58	1	0	0
3/12/2023 2:00	1	1	0	74	63	10	9	156	0	1	0
3/12/2023 3:00	0	0	0	116	143	26	36	321	0	0	0
3/12/2023 4:00	0	0	0	278	260	59	50	647	0	0	0
3/12/2023 5:00	0	0	0	565	546	327	123	1561	0	0	0

Flight data with terminal location added to the dataset.

Step 2:

The existing dataset was used to create new datasets with the car entering the lot offset by 1 to 4 hours. This was done as it is likely that most passengers will arrive at the airport several hours before their flight, and the amount of cars entering the lot in a specific hour is likely dependent of the number of flights in the next several hours.

Four datasets were created with the cars entering the lot offset (1 and 2 hour offset shown):

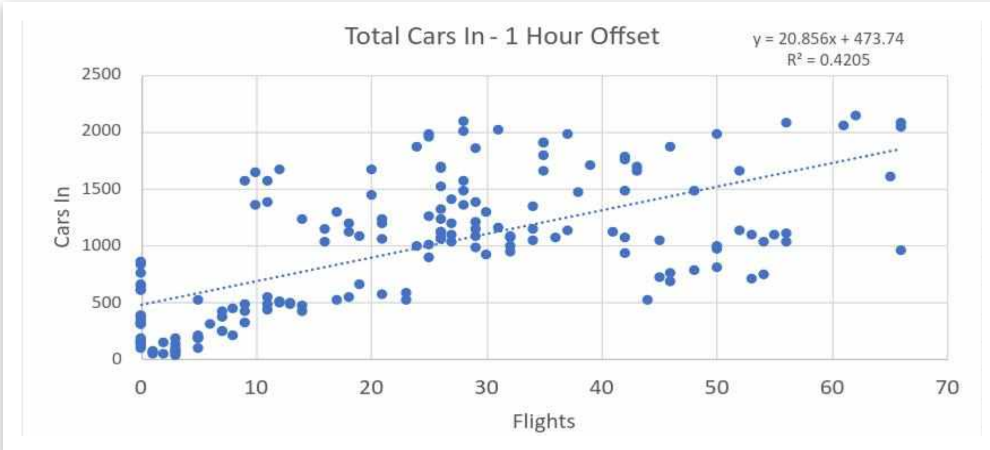
Time	Flights	Flights T1	Flights T2	T1 Short Term	T1 Long Term	T2 Short Term	T2 Long Term	Total Cars In	Europe	South Asia	Middle East
3/12/2023 0:00	7	7	0						0	2	4
3/12/2023 1:00	5	4	1	47	25	18	8	98	1	0	0
3/12/2023 2:00	1	1	0	24	24	6	4	58	0	1	0
3/12/2023 3:00	0	0	0	74	63	10	9	156	0	0	0
3/12/2023 4:00	0	0	0	116	143	26	36	321	0	0	0
3/12/2023 5:00	0	0	0	278	260	59	50	647	0	0	0

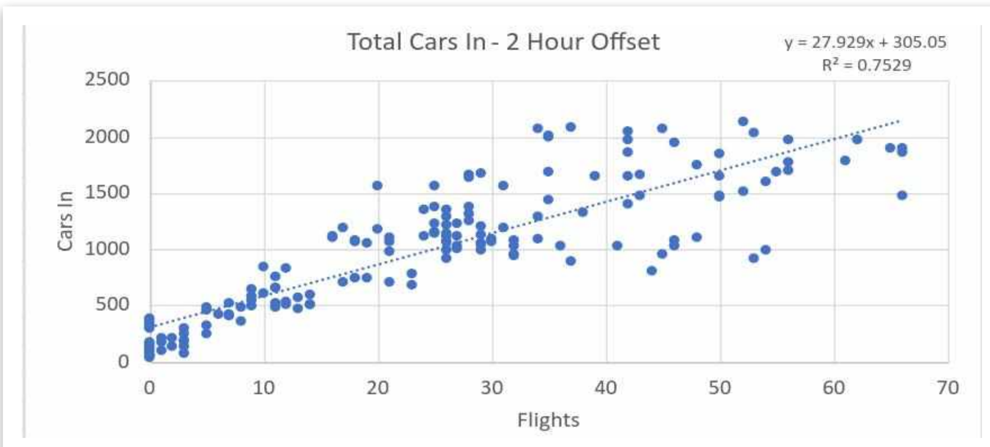
The quantity of cars entering the lot offset by 1 hour.

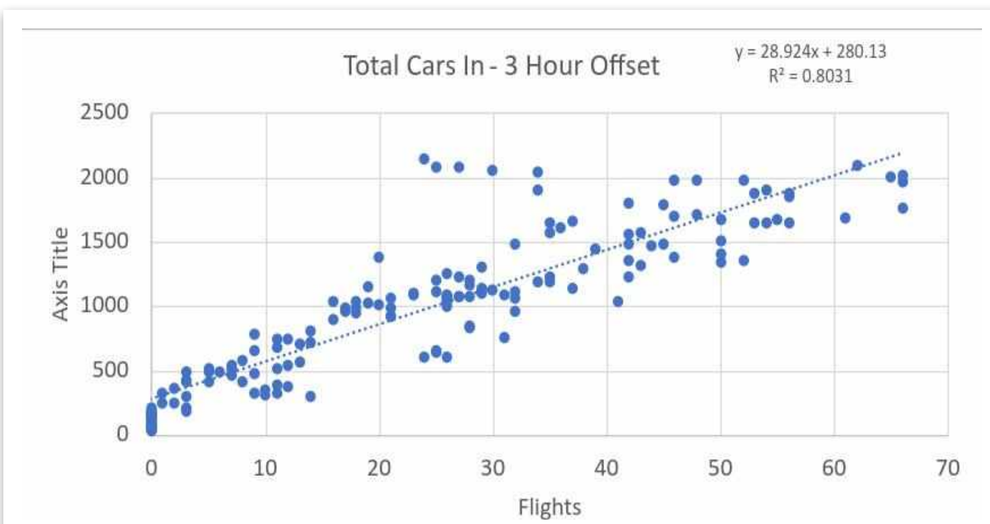
Time	Flights	Flights T1	Flights T2	T1 Short Term	T1 Long Term	T2 Short Term	T2 Long Term	Total Cars In	Europe	South Asia	Middle East
3/12/2023 0:00	7	7	0						0	2	4
3/12/2023 1:00	5	4	1						1	0	0
3/12/2023 2:00	1	1	0	47	25	18	8	98	0	1	0
3/12/2023 3:00	0	0	0	24	24	6	4	58	0	0	0
3/12/2023 4:00	0	0	0	74	63	10	9	156	0	0	0
3/12/2023 5:00	0	0	0	116	143	26	36	321	0	0	0

The quantity of cars entering the lot offset by 2 hours.

In addition, to best determine which dataset to use, a simple linear regression was created for all models. Looking at the data, it appears that the number of cars entering the lot best correlates to the number of flights in 3 hours. Using this dataset with the cars offset by three hours will likely provide the best starting point for further analysis to be performed in the subsequent assignments.


Total Cars In - 1 Hour Offset
y = 20.856x + 473.74
R² = 0.4205
2500
2000
1500
1000
500
0
Cars In
0
10
20
30
40
50
60
70
Flights


Total Cars In - 2 Hour Offset
y = 27.929x + 305.05
R² = 0.7529
2500
2000
1500
1000
500
0
Cars In
0
10
20
30
40
50
60
70
Flights


Total Cars In - 3 Hour Offset
y = 28.924x + 280.13
R² = 0.8031
2500
2000
1500
1000
500
0
Axis Title
0
10
20
30
40
50
60
70
Flights

Multi Linear Regression Modeling

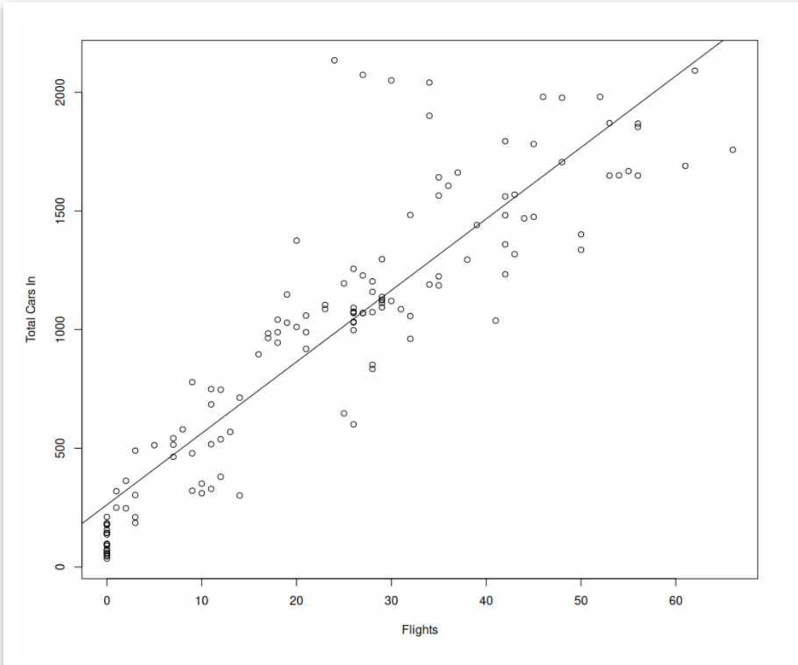
- Linear regression was used to determine the quantity of cars based on number of flights. Multiple linear regression models were created to see if individual variables could be used to generate a model that would accurately predict the number of cars in various situations. Data for the linear regressions was split into 80% for training and 20% for testing. Each model was evaluated to calculate the error and determine if the model was accurate enough to be used for predicting cars entering the different lots. Based on the results of the data cleaning step, the flight data was offset 3 hours from the car data, such that the number of cars is calculated by the number of flights that are departing 3 hours after the cars arrive.

  Six models were generated using linear regression.

  - Model 1: Total Flights vs Total Cars In
  - Model 2: Flights Departing from T1 vs Cars Parking in T1
  - Model 3: Flights Departing from T2 vs Cars Parking in T2
  - Model 4: Regions vs Total Cars In
  - Model 5: Regions vs Short-Term Parking
  - Model 6: Regions vs Long-Term Parking

Model 1: Total Flights vs Total Cars In

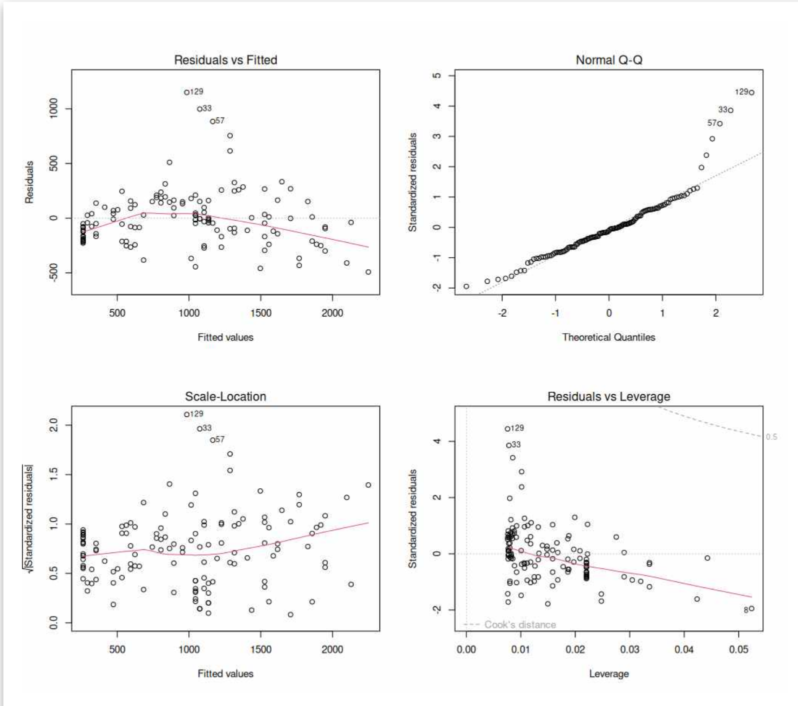
The first model used a simple combination of total flights and total cars.



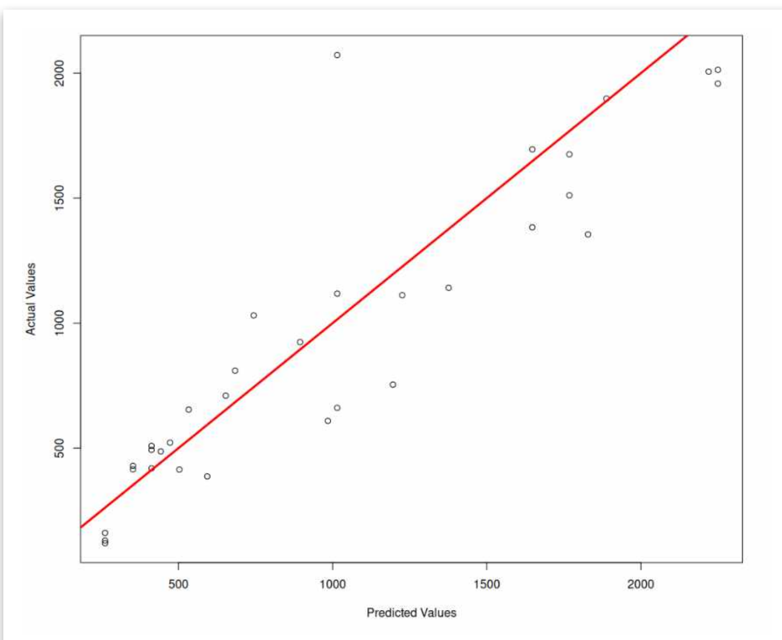

```
Call:
lm(formula = Total.Cars.In ~ Flights, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-491.95 -166.55  -27.45  138.41 1150.09 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  262.031     38.568   6.794 3.54e-10 ***
Flights       30.120      1.306  23.054  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 259.8 on 130 degrees of freedom
Multiple R-squared:  0.8035,	Adjusted R-squared:  0.802 
F-statistic: 531.5 on 1 and 130 DF,  p-value: < 2.2e-16
```

- The first model generated an equation with a flight coefficient of 30.12 and an intercept of 262.03. The residual standard error was 259.8 and the R^2 value was 0.80. The p-value for the flights variable and the overall equation were both nearly 0 and the F-statistic is far above 1, indicating that the variables are statistically significant and that there is a linear relationship between flights and total cars in. The R^2 value of 0.80 indicates that the model also has a good fit.

  To further check the validity of the regression model, the residual and Q-Q plots were generated.



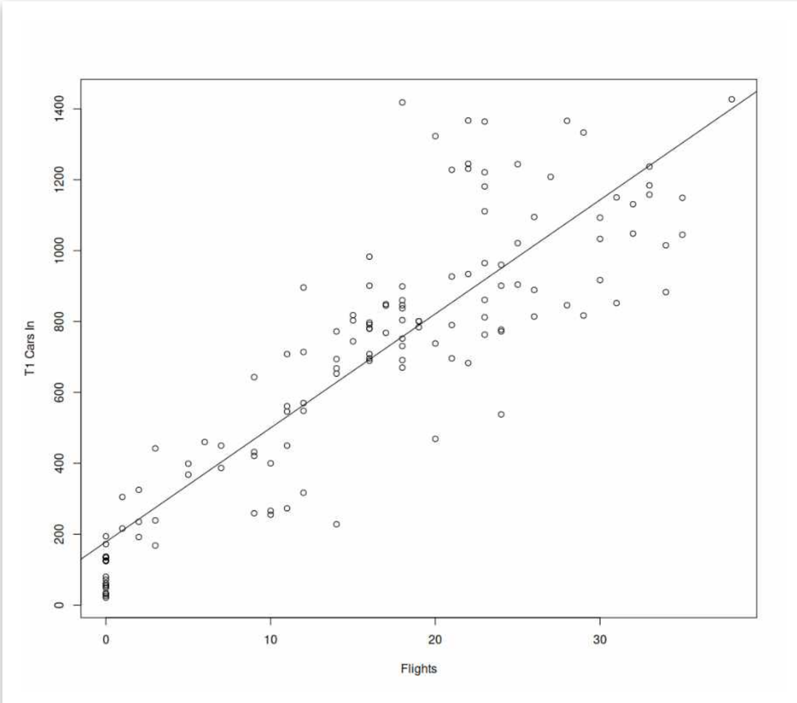
- ➢ The residuals vs fitted chart shows a line that increases and then decreases, indicating that there is not evidence of homoscedasticity, indicating that there are some outliers in the data. This can also be seen with the Q-Q plot, where most of the data is normally distributed, although the chart deviates on the right side, indicating the presence of some outliers.
- ➢ The model was tested with the training data. Predicted vs actual values are shown in the chart below. The test data had a residual standard error of 281.78 and an R2 value of 0.80, indicating performance was similar to the training data.



- Most of the predicted values are close to actual values. There is a tendency that when the quantity of cars is overestimated, the estimates are further from the actual value than compared to when the number of cars is underestimated.

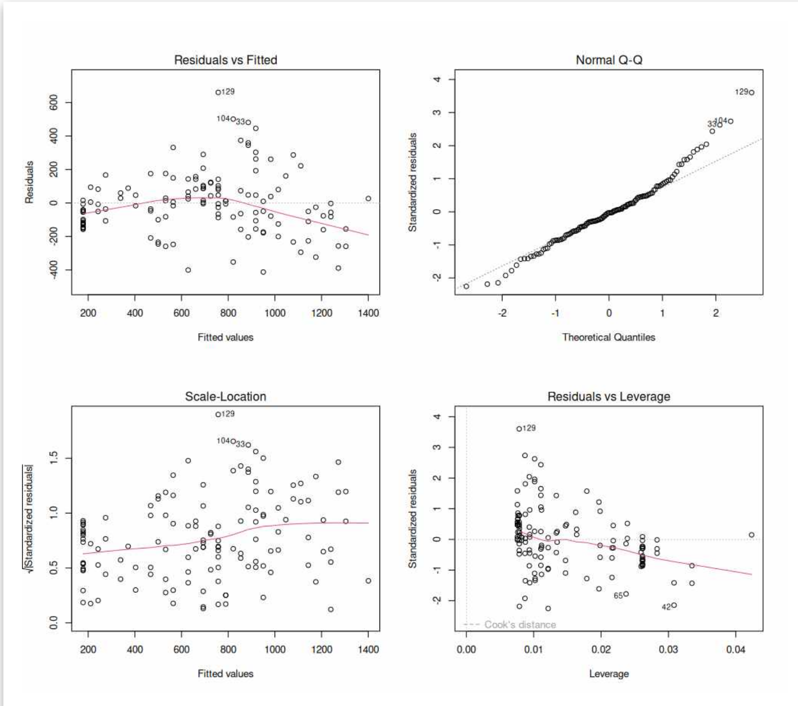
Model 2: Flights Departing from T1 vs Cars Parking in T1

The second model used data with flights departing from T1 and cars parking in T1.

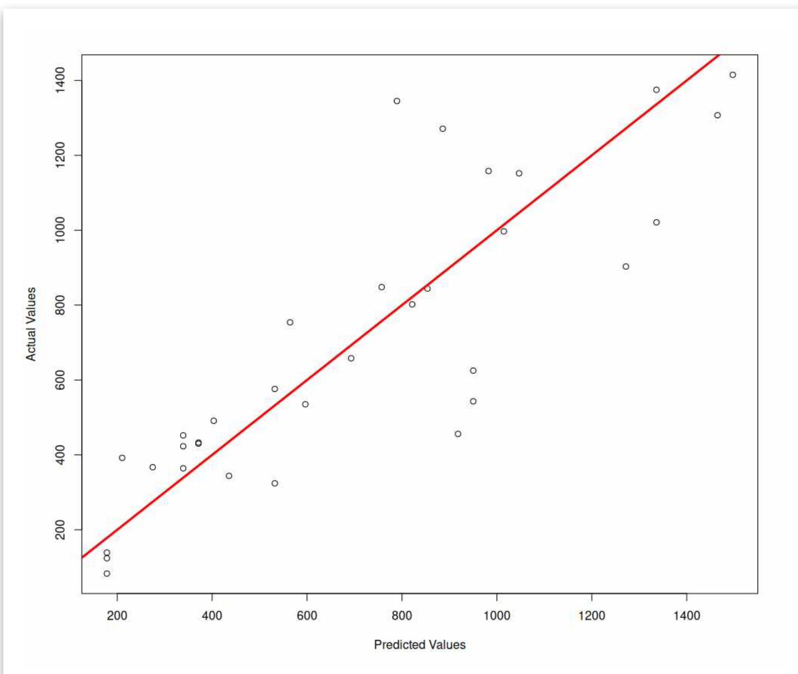



```
Call:
lm(formula = T1.Total.Cars ~ Flights.T1, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-412.21 -107.60   -5.81   87.53  660.78 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  178.229     29.764   5.988 1.95e-08 ***
Flights.T1    32.166      1.563  20.575  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 184.1 on 130 degrees of freedom
Multiple R-squared:  0.7651,	Adjusted R-squared:  0.7633 
F-statistic: 423.3 on 1 and 130 DF,  p-value: < 2.2e-16
```

- The flights from T1 vs cars in T1 model generated an equation with a flight coefficient of 32.17 and an intercept of 178.23. The residual standard error was 184.1 and the R2 value was 0.77. Similar to the first model, the p-value for the flights variable and the overall equation were both nearly 0 and the F-statistic is far above 1, indicating that the variables are statistically significant and that there is a linear relationship between flights and cars in T1. The R2 value of 0.77 indicates that the model also has a good fit.



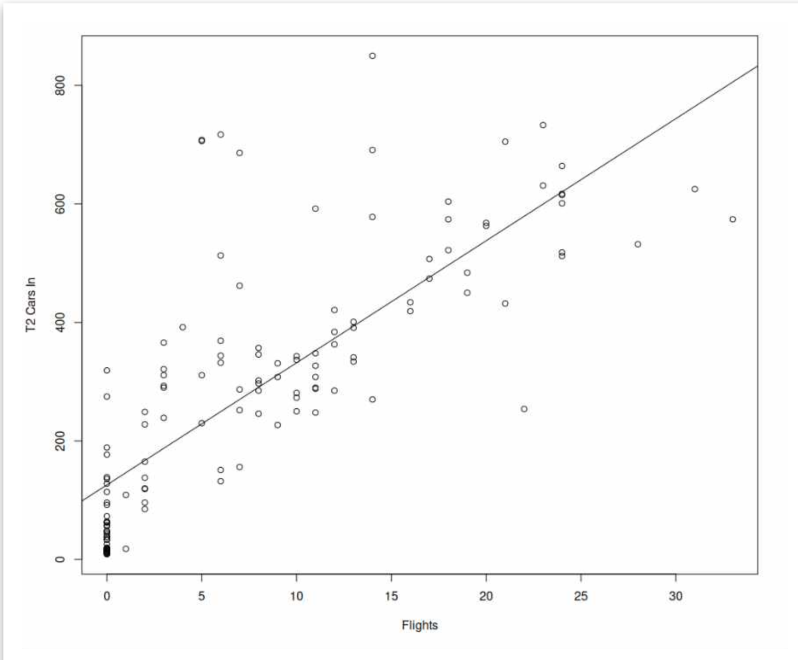
➢ The residual and Q-Q charts look similar to the first model, some outliers are present in the data. The test data had a residual standard error of 216.02 and an R2 value of 0.87, indicating performance was similar to the training data.



➢ The predicted vs actual graph is similar to the first model.

Model 3: Flights Departing from T2 vs Cars Parking in T2

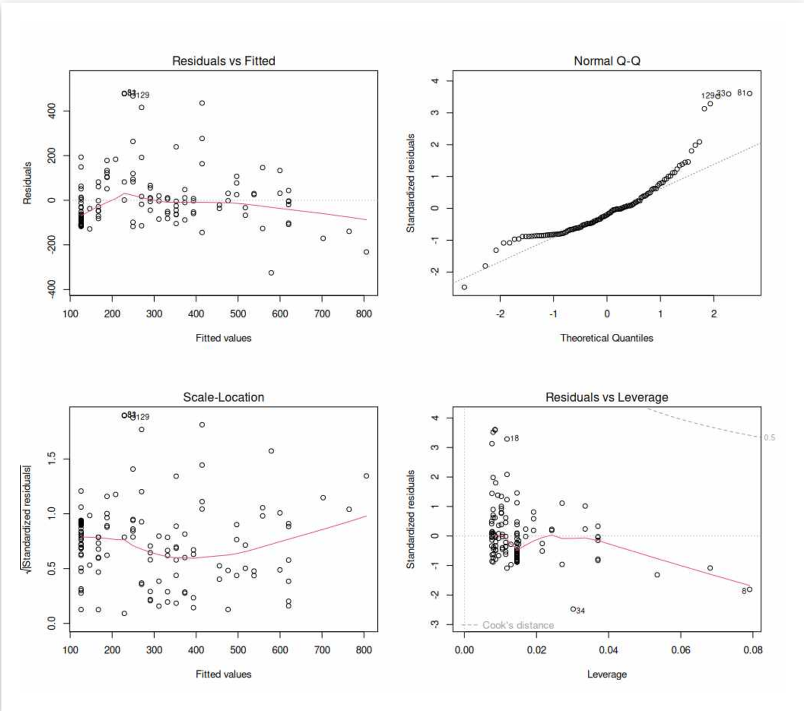
The third model used data with flights departing from T2 and cars parking in T2.



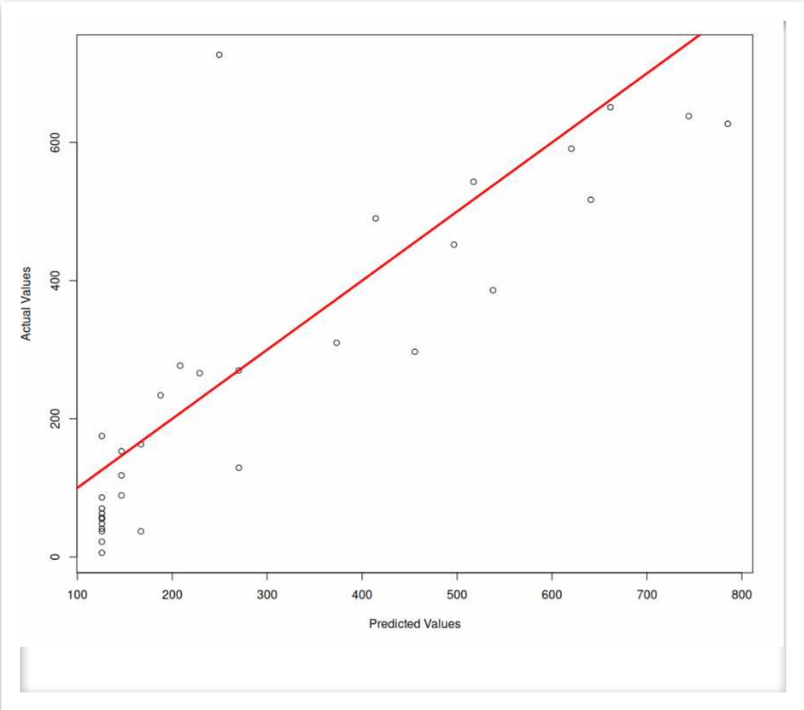
- The flights from T2 vs cars in T2 model generated an equation with a flight coefficient of 20.60 and an intercept of 125.88. The residual standard error was 210.6 and the R2 value was 0.62, slightly lower than the previous models. Similar to previous models, the p-value for the flights variable and the overall equation were both nearly 0 and the F-statistic is far above 1, indicating that the variables are statistically significant and that there is a linear relationship between flights departing T2 and cars in T2.

```
Call:
lm(formula = T2.Total.Cars ~ Flights.T2, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-325.15  -87.94  -27.30   48.69  479.10 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   125.88      16.11   7.811 1.67e-12 ***
Flights.T2     20.60       1.42  14.513  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133.4 on 130 degrees of freedom
Multiple R-squared:  0.6183,    Adjusted R-squared:  0.6154 
F-statistic: 210.6 on 1 and 130 DF,  p-value: < 2.2e-16
```

➢ The residual and Q-Q charts look similar to the previous model, with the Q-Q chart showing the data is less normality compared to the T1 model. The test data had a residual standard error of 121.50 and an R2 value of 0.62, indicating performance was similar to the training data.



➢ The predicted vs actual graph is similar to the previous models.

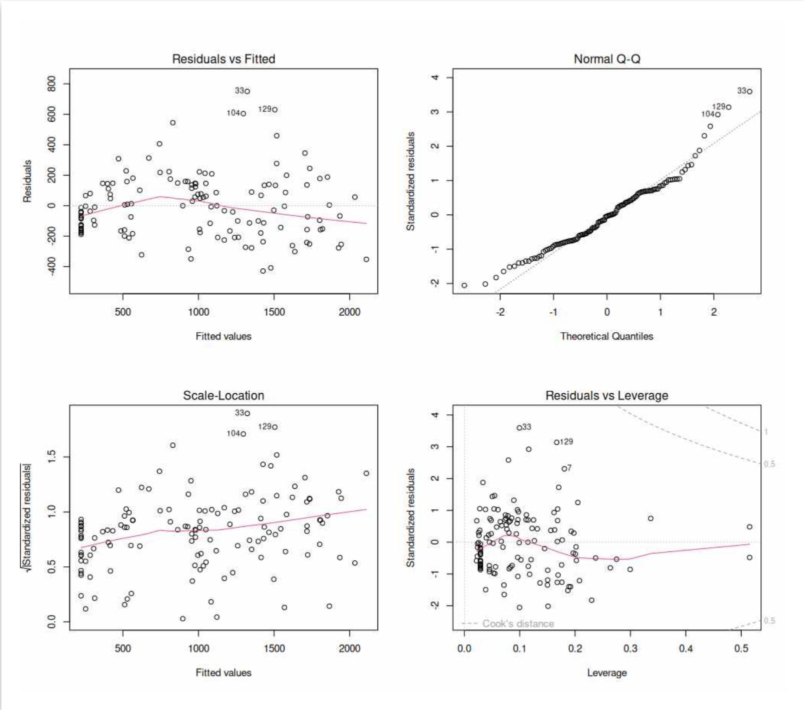


Model 4: Regions vs Total Cars In

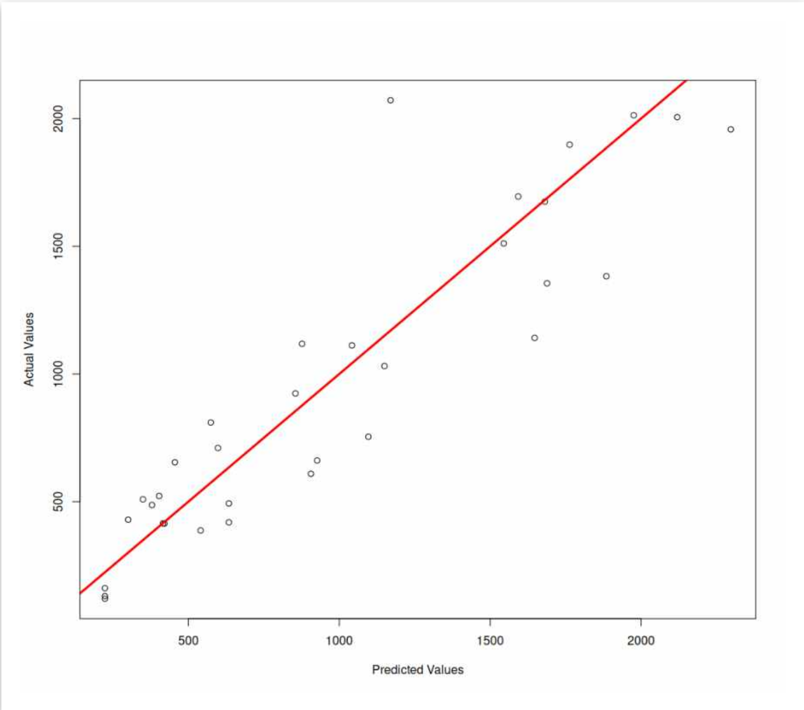
The fourth model broke up the flights into regions and looked at predicting the total amount of cars in the lots.

```
Call:
lm(formula = Total.Cars.In ~ Europe + South.Asia + Middle.East +
    Africa + Vietnam + CentralAsia + Japan + NearAsia + China +
    PacificOcean + America + Aus + Korea, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-428.81 -157.30   -8.61  137.24  750.34

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   223.217     37.522   5.949 2.83e-08 ***
Europe         23.575     13.487   1.748 0.083065 .
South.Asia     29.816      6.292   4.739 6.06e-06 ***
Middle.East    24.118     11.867   2.032 0.044366 *
Africa        132.751    160.385   0.828 0.409511
Vietnam        30.025      6.857   4.379 2.60e-05 ***
CentralAsia   160.369     19.451   8.245 2.63e-13 ***
Japan          26.756      3.088   8.664 2.81e-14 ***
NearAsia       16.320     11.144   1.464 0.145717
China          59.822     17.518   3.415 0.000875 ***
PacificOcean   56.513     16.586   3.407 0.000898 ***
America        28.155      7.047   3.995 0.000113 ***
Aus            30.710     19.873   1.545 0.124943
Korea           4.712     22.839   0.206 0.836913
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 220 on 118 degrees of freedom
Multiple R-squared:  0.872,     Adjusted R-squared:  0.8579
F-statistic: 61.84 on 13 and 118 DF,  p-value: < 2.2e-16
```

- Since this model uses multiple input variables, a chart of the linear regression cannot be shown. The coefficients for each region were generated in the linear regression and are shown above. Most regions have a p-value that is significant. South Asia, Vietnam, Central Asia, Japan, China, Pacific Ocean, and America all have p-values near 0. The Middle East has a p-value below 0.05 and the p-value for Europe is below 0.1. Both regions could be considered significant, but not to the same degree as the variables with p-values near 0. Africa, Near Asia, Australia, and Korea do not have significant p-values, which is likely a result of those locations having fewer flights in the dataset compared to other regions.
- The coefficients can be used to compare the cars per flight to different regions. For example, most regions typically have 20-30 cars per flight, while China and the Pacific Ocean are higher at almost 60 cars per flight. Central Asia has the highest at 160 cars per flight.
- The overall model shows a linear relationship, with the F-statistic above 1 and the p-value of the model near 0. The R2 value is 0.872, indicating that the model is a good fit.



- ➢ The residuals and Q-Q charts again look similar to other models. The test data had a residual standard error of 343.31 and an R2 value of 0.87. This indicates a good fit for the model, with a slightly higher root square error for the testing data compared to the training data.
- ➢ The predicted vs actual graph is similar to the previous models.



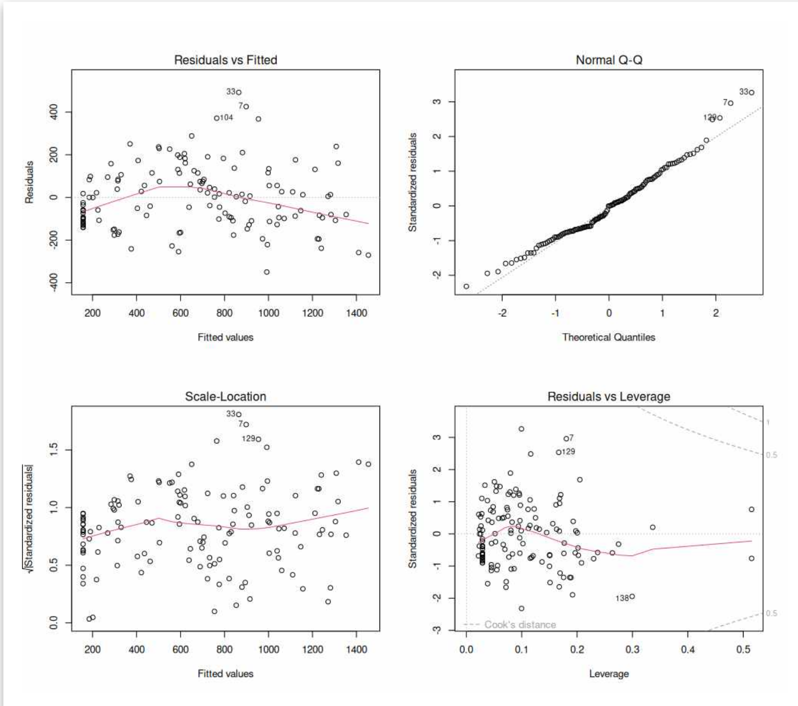
- ➢ Since the model shows a statistically significant linear relationship between most of the variables, we can use the coefficients to predict the number of cars that will be parking for flights to a specific region. The coefficients for all regions except Africa, Australia, and Korea which had non-significant p-values, could be used for calculating total cars parking based on the number of flights in those regions.

Model 5: Regions vs Short-Term Parking

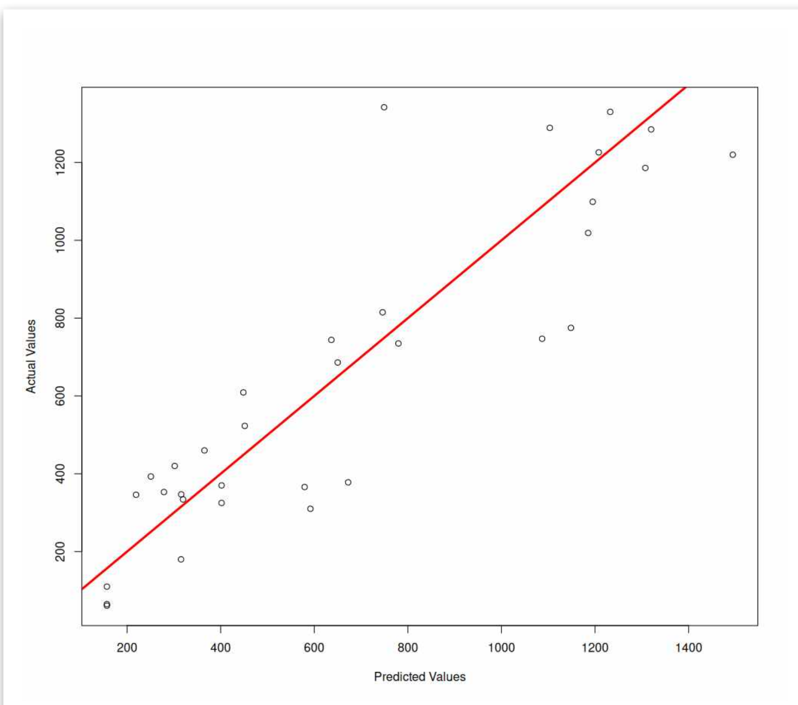
The fifth model broke up the flights into regions and looked at predicting the total amount of cars in the short-term lots.

```
Call:
lm(formula = Short.Term ~ Europe + South.Asia + Middle.East +
    Africa + Vietnam + CentralAsia + Japan + NearAsia + China +
    PacificOcean + America + Aus + Korea, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-349.43 -107.35   -0.09   95.96  492.13

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  156.8862    27.0809   5.793 5.85e-08 ***
Europe        22.9036     9.7339   2.353 0.020279 *
South.Asia    27.9357     4.5414   6.151 1.08e-08 ***
Middle.East   16.5153     8.5651   1.928 0.056232 .
Africa       125.7236   115.7554   1.086 0.279642
Vietnam       16.5363     4.9489   3.341 0.001117 **
CentralAsia   97.7351    14.0386   6.962 2.03e-10 ***
Japan         10.6733     2.2289   4.789 4.92e-06 ***
NearAsia      19.4333     8.0429   2.416 0.017216 *
China         44.3152    12.6432   3.505 0.000646 ***
PacificOcean  28.4807    11.9707   2.379 0.018954 *
America       22.1088     5.0860   4.347 2.94e-05 ***
Aus           33.7191    14.3430   2.351 0.020387 *
Korea         -0.5609    16.4841  -0.034 0.972911
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 158.8 on 118 degrees of freedom
Multiple R-squared:  0.857,     Adjusted R-squared:  0.8412
F-statistic: 54.39 on 13 and 118 DF,  p-value: < 2.2e-16
```

- Similar to the model with all cars, most regions have a p-value that is significant. South Asia, Vietnam, Central Asia, Japan, China, Pacific Ocean, and America all again have p-values near 0. The Middle East no longer has a p-value below 0.05 but the value is still below 0.1. The p-value for Europe is now below 0.05. Both regions could be considered significant, but not to the same degree as the variables with p-values near 0. Africa and Korea again do not have significant p-values. Near Asia and Australia have p-values that can be considered significant, unlike in the previous model. It may be that due to the close proximity of these countries to Korea that the passengers likely park primarily in the short-term lots, which skews the distribution but results in a statistically significant coefficient in the linear model when only looking at short term parking.
- The F-statistic is above 1 and the p-value for the model is near 0, indicating a good fit.



➢ The residual and Q-Q graphs are again similar to the other models, with the residual square error 246.22 and the R2 value 0.86 for the test model. This indicates that model 5 is also a good fitting model.



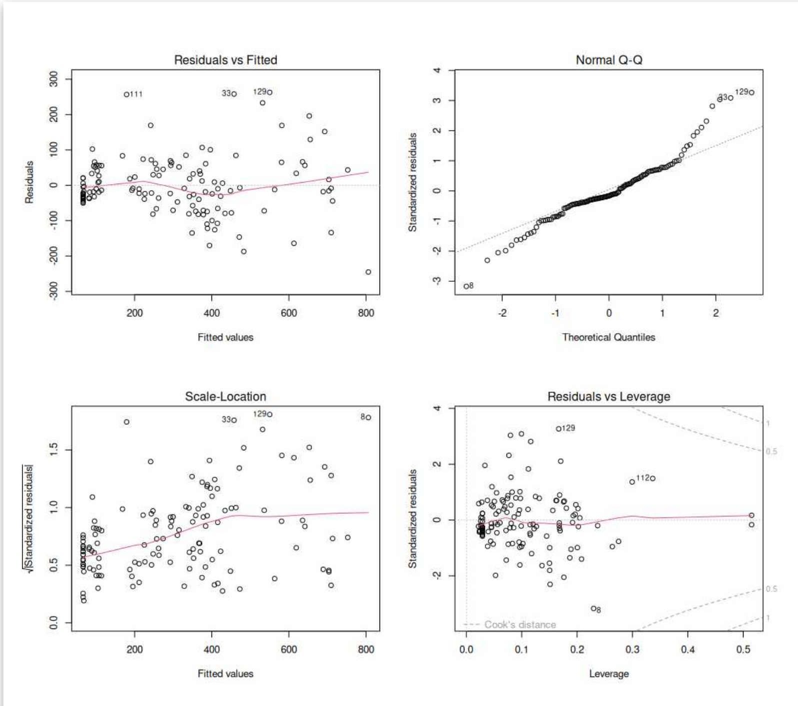
➢ The predicted vs actual value chart is similar to the other models.

Model 6: Regions vs Long-Term Parking

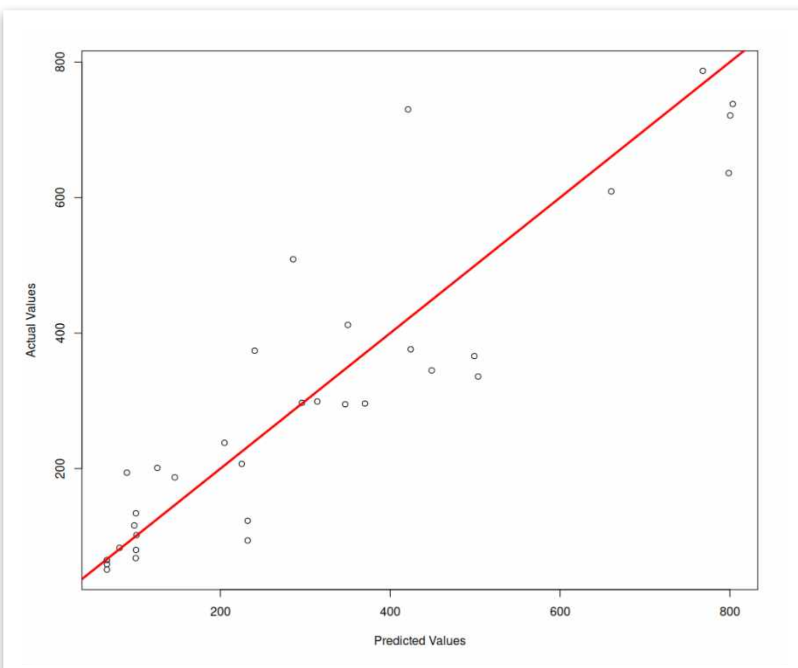
The final sixth model broke up the flights into regions and looked at predicting the total amount of cars in the long-term lots.

```
Call:
lm(formula = Long.Term ~ Europe + South.Asia + Middle.East +
    Africa + Vietnam + CentralAsia + Japan + NearAsia + China +
    PacificOcean + America + Aus + Korea, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-245.18  -38.52  -12.99   45.26  262.70

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   66.3305    15.0291   4.413 2.26e-05 ***
Europe         0.6713     5.4020   0.124   0.9013
South.Asia     1.8807     2.5203   0.746   0.4570
Middle.East    7.6031     4.7534   1.600   0.1124
Africa         7.0269    64.2411   0.109   0.9131
Vietnam       13.4889     2.7465   4.911 2.94e-06 ***
CentralAsia   62.6343     7.7910   8.039 7.81e-13 ***
Japan         16.0831     1.2370  13.002  < 2e-16 ***
NearAsia      -3.1131     4.4636  -0.697   0.4869
China         15.5064     7.0166   2.210   0.0290 *
PacificOcean  28.0326     6.6434   4.220 4.83e-05 ***
America        6.0463     2.8226   2.142   0.0342 *
Aus           -3.0087     7.9600  -0.378   0.7061
Korea          5.2727     9.1482   0.576   0.5655
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 88.13 on 118 degrees of freedom
Multiple R-squared:  0.8489,    Adjusted R-squared:  0.8323
F-statistic:    51 on 13 and 118 DF,  p-value: < 2.2e-16
```

- In this model, many of the coefficients are no longer statistically significant. Only Vietnam, Central Asia, Japan, and the Pacific Ocean have p-values near 0. China and America have p-values below 0.05. The remaining variables do not have statistically significant p-values for the coefficients.
- While the F-statistic and the p-value for the model still indicate that a linear relationship exists, since many of the p-values are not statistically significant, they cannot be used to accurately predict the number of cars that will park in the lots for those regions. Since there are more cars that park in the short-term lots than long-term, there is likely a lack of data that is resulting in poor fits for the coefficients of some regions.



- ➢ The residual vs fitted graph has the closest to a straight line of all the models, indicating the model is close to homoscedastic, however the ends of the Q-Q plot deviate the most from the center, indicating that the data for this model is the least normally distributed of all models. The residual square error 131.40 and the R2 value 0.85 for the test model which still indicates that model 6 appears to be a good fit, although it does not satisfy all the conditions needed for a linear regression.



- Even though not all coefficients are statistically significant, the overall model does show a good fit at predicting the amount of cars parking in the long-term lots.

Linear Analysis Conclusion

Overall, all of the models had a good fit and performed well with the test data. The different models can be used to predict the number of cars based on the number of flights. The first three models can be used to predict the number of cars with just the total number of flights for each terminal or just the number of total flights together. If a more specific model is needed, the second three models can be used to predict the number of cars that will use the lots for a flight to a specific region. Central Asia consistently showed more cars per flight in the three regional models compared to other regions.

Project Conclusion

The linear model proved to be the most effective model for determining the number of cars per flight. The first three models can be used to calculate overall cars in the lots, cars in T1 parking, and cars in T2 parking. The second three models can be used to predict the number of cars per flight depending on the region of the flight. Of the three models, model 4 which predicted the overall number of cars had the most statistically significant coefficients, making it the most accurate if just a simple estimate is needed for a flight to a specific region.

Our initial questions for the project were:

- **Is there any correlation between the number of flights and the number of private cars parked at the airport?**

- **Which flight or destination (Europe, America, Asian, etc.) has the most significant impact on traffic in the parking area?**

We can answer question 1 by using the first three linear regression models. Overall for every flight, there is an average of 30.12 cars that will park at the airport. The second question can be answered by the second three models, specifically using model 4 if just a general number of cars is needed. Using the coefficients for model 4, Central Asia has the most significant effect on the amount of cars parking at the airport, with 160.37 cars per flight, far exceeding the number of cars for flights to other regions.

References

Incheon International Airport Corporation. (2023). *Departure Schedule.* Retrieved March 17, 2023, from https://www.airport.kr/ap/en/dep/depPasSchList.do

Rebecca Bevans. (2020). Multi Linear Regression/ A Quick Guide (Examples) Retrieved April 11, 2023 from https://www.scribbr.com/statistics/multiple-linear-regression/