Problem Solving Task 1

MXN500

Due date	10/05/20 11:59pm
Last name	Nayak
First name	Drashti
Student number	n10599568

About this Problem Solving Task

On completion of this unit you should be able to analyse a data set relating to one of the natural sciences, business, or IT, estimate parameters in an empirical relationship and make inferences based on model diagnostics.

Here's a checklist of the steps you need to go through before submitting this online.

- **Individualise your data** by subsetting to keep only the rows with your student number in it.
- **Answer the questions** in the relevant areas in this document.
  - For all steps that require you to show your working, give enough detail (and explanatory text where needed) that anyone reading can follow the logic of your solution.
  - Correct answers with no working will not attract full marks.
  - Provide any R code used to solve the problem in your R script file
- **Interpret the analysis** in the space provided in the Interpretation section
  - Ensure all plots are included at a resolution that makes them readable and that they have either a caption or a title
- **Provide your completed document** by uploading it to Blackboard as either a Microsoft Word file of format .doc or .docx or a PDF file of format .pdf. Pages files will be awarded a mark of 0
- **Provide your R script** file
  - PSTs with no script files will be considered incomplete and awarded a mark of 0.
  - The script file should
    - have file extension `.R`
    - contain all code required to read in the downloaded data, and
    - perform all the steps of the analysis that require the use of R

Submit your PST and R script file on Blackboard by the due date. Ensure you have attached both files.

This is individual assessment. You should not work with others to complete the task, share your work, or ask other students for their work.

Your solution must be your own. You are not permitted to copy, summarise, or paraphrase the work of others in your solution.

Please ensure you are familiar with QUT's policy on Academic Integrity

Introduction

The United States Department of Transportation publishes detailed flight data. This data is available via a range of websites, however you will use data available on the Bureau of Transportation Statistics website.

Here, we are particularly interested in understanding if the airline carrier has a major impact on the delay in departure and arrival times. We are also interested in understanding whether airlines are able to make up time on their flight if they had a departure delay.

The full dataset is very large. As such, we will only be considering a subset of this data from January 2020.

Set up

Download `PST1Random.csv` from Blackboard. Read this into `R` and filter by your student number. The Origin listed will indicate the origin that you will be investigating.

Download `PST1Data.csv` from Blackboard. Read this into `R` and filter by your unique Origin. Since you have the same Origin for all entries in your data frame, remove both Origin and OriginStateName.

Download `PST1AirportCodes.csv` from Blackboard. Read this into `R`. Since there are no airport names in the original dataset, use a join to add the destination airport name to your dataframe.

Download `PST1AirlineCodes.csv` from Blackboard. Read this into `R`. Since there are no airline names in the original dataset, use a join to add the reporting airline name to your dataframe.

Section A - Data and Visualisation (30%)

A1 Data Structure (10%)

Before we can start fitting models, we need to get a clear picture of our data.

Exercise: Create a data dictionary of the variables in your data frame.

Answer:

Filter

	Variable	description	type	class
1	FlightDate	date when flight departs	character	character
2	Reporting_airline	name of airline carrier	character	character
3	Destination	Codename Of destination	character	character
4	DestStateName	destination state name of USA	character	character
5	DepDelay	Departure Delay count	double	numeric
6	ArrDelay	Arrival Delay count	double	numeric
7	Airport Name	name of the airport it arrived	character	character
8	Airline name	name of the airline	character	character

Exercise: Currently, many of your columns are classed as characters. Convert your data to more appropriate classes. List your conversions below.

Answer:

	Variable	from	to
1	FlightDate	character	Date
2	Reporting_airline	character	factor
3	Destination	character	factor
4	DestStateName	character	factor
5	DepDelay	double	numeric
6	ArrDelay	double	numeric
7	Airport Name	character	factor
8	Airline name	character	factor

Exercise: Create a table below, summarising the number of observations, mean, median, and standard deviation of departure delays.

Answer:

Table of summaries of Departure delays

No.of observations	1662
No. of missing values	07
Mean	6.036
Median	-4
Standard Deviation	40.58908

Exercise: Repeat the above exercise, reporting the data for each of the five most popular airlines.

Answer:

Mean of top 5 Airlines:

Alaska Airlines	3.52
Allegiant Air LLC	4.96
American Airlines	13.1
Delta Air Lines	-0.662
Envoy Air Inc.	3.05

Median of top 5 Airlines:

Alaska Airlines	-4
Allegiant Air LLC	0
American Airlines	-2
Delta Air Lines	-3
Envoy Air Inc.	-3

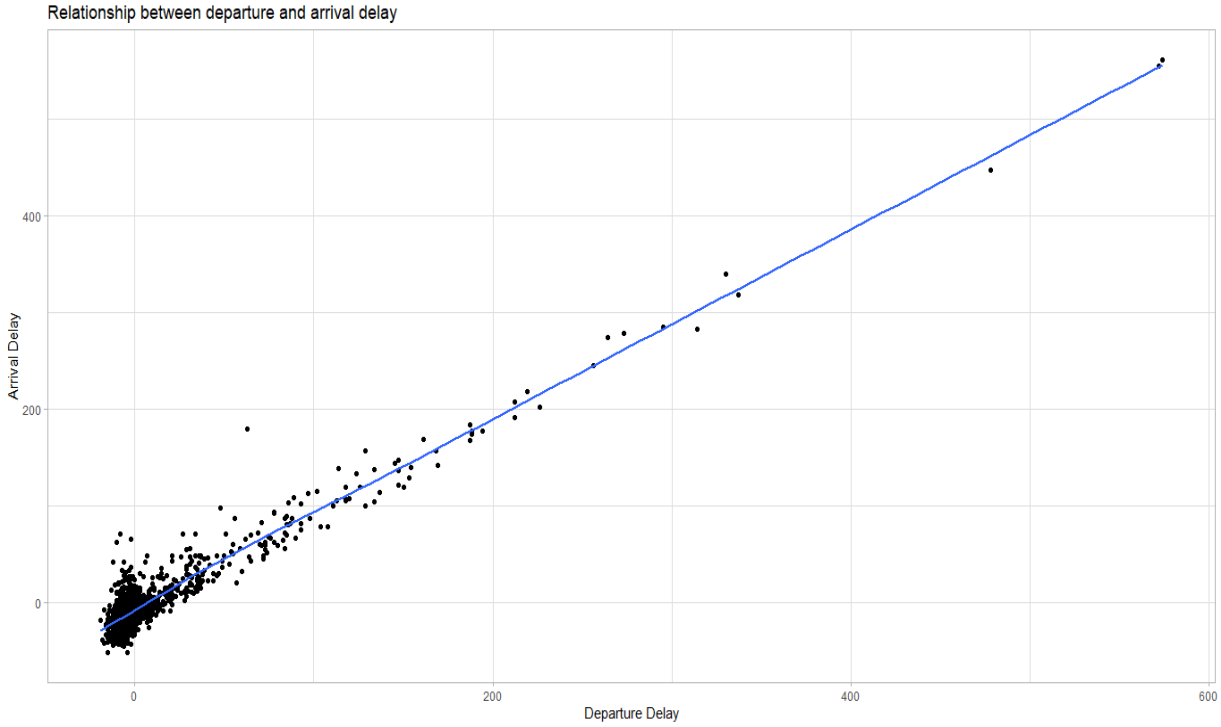
Standard Deviation of top 5 Airlines:

Alaska Airlines	24.7
Allegiant Air LLC	16.7
American Airlines	
Delta Air Lines	11.1
Envoy Air Inc.	20.8

A2 Graphical Summaries (15%)

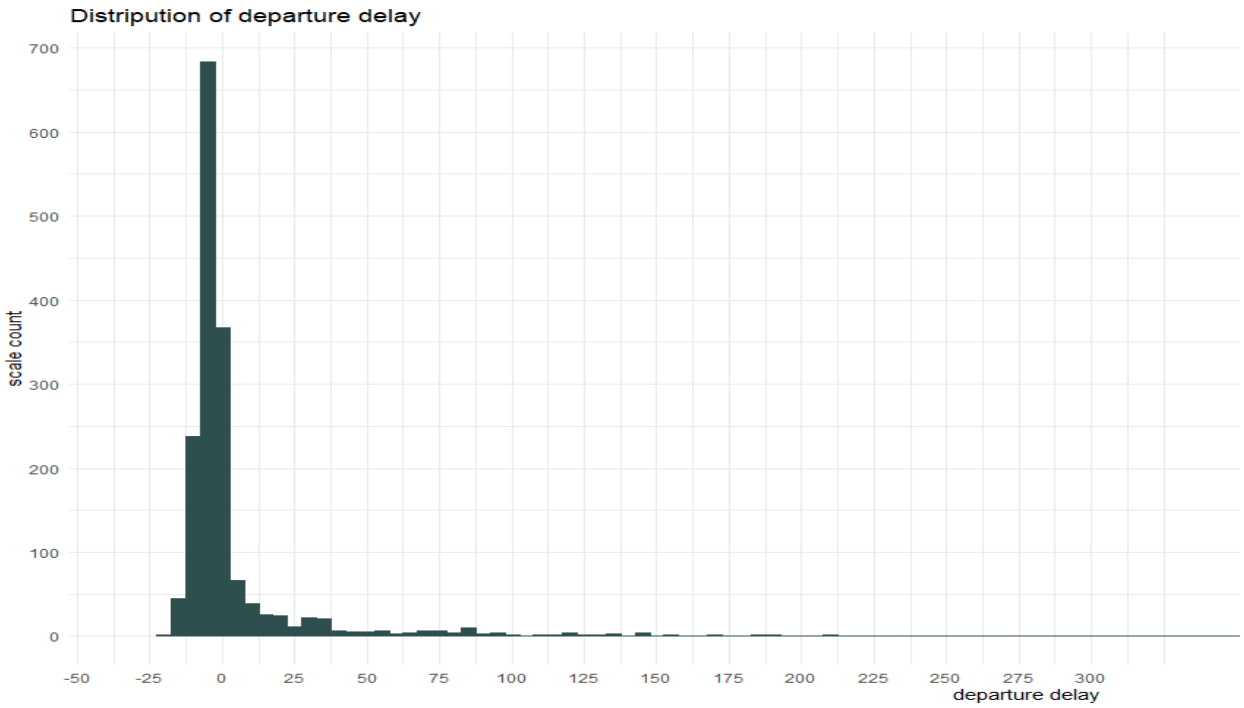
Exercise: Create a graphically excellent plot that shows the relationship between departure delay and arrival delay.

Answer:



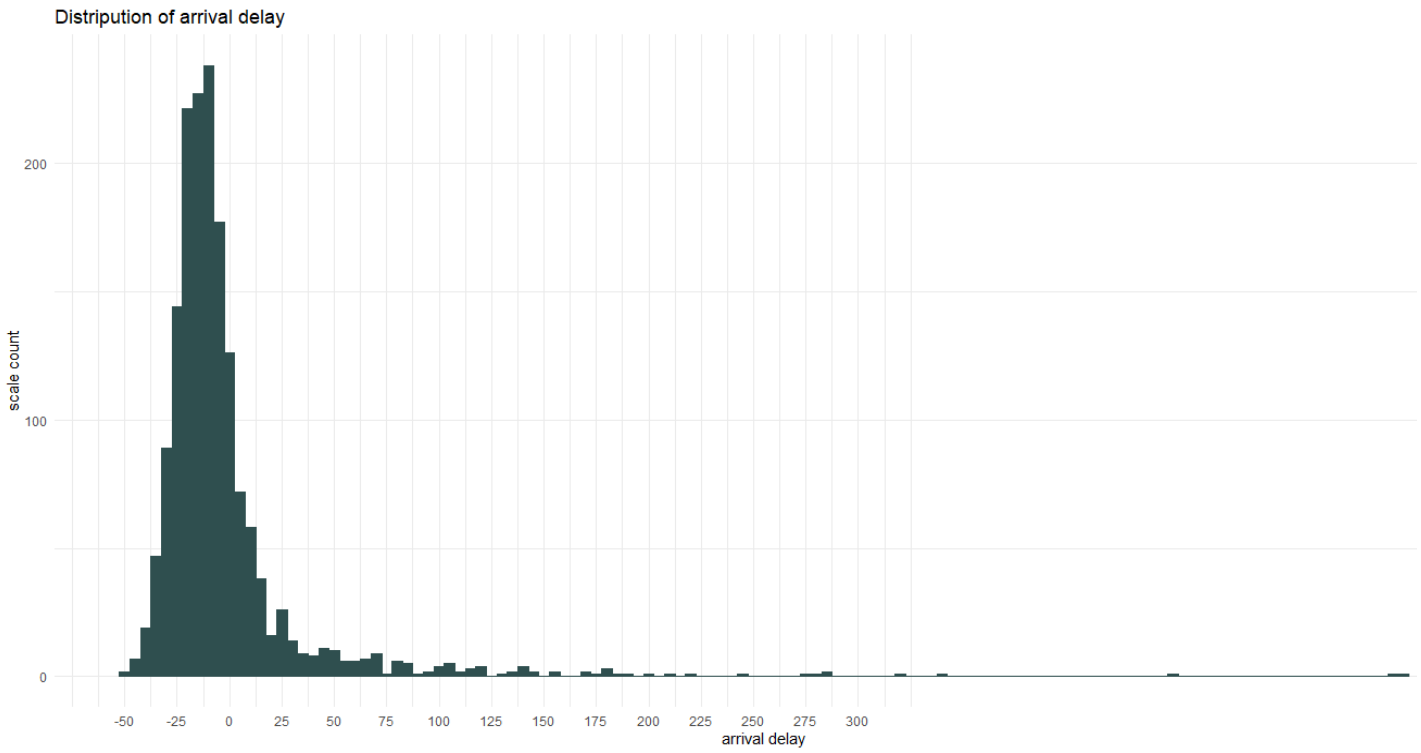
Exercise: Create a graphically excellent plot that shows the distribution of departure delays.

Answer:



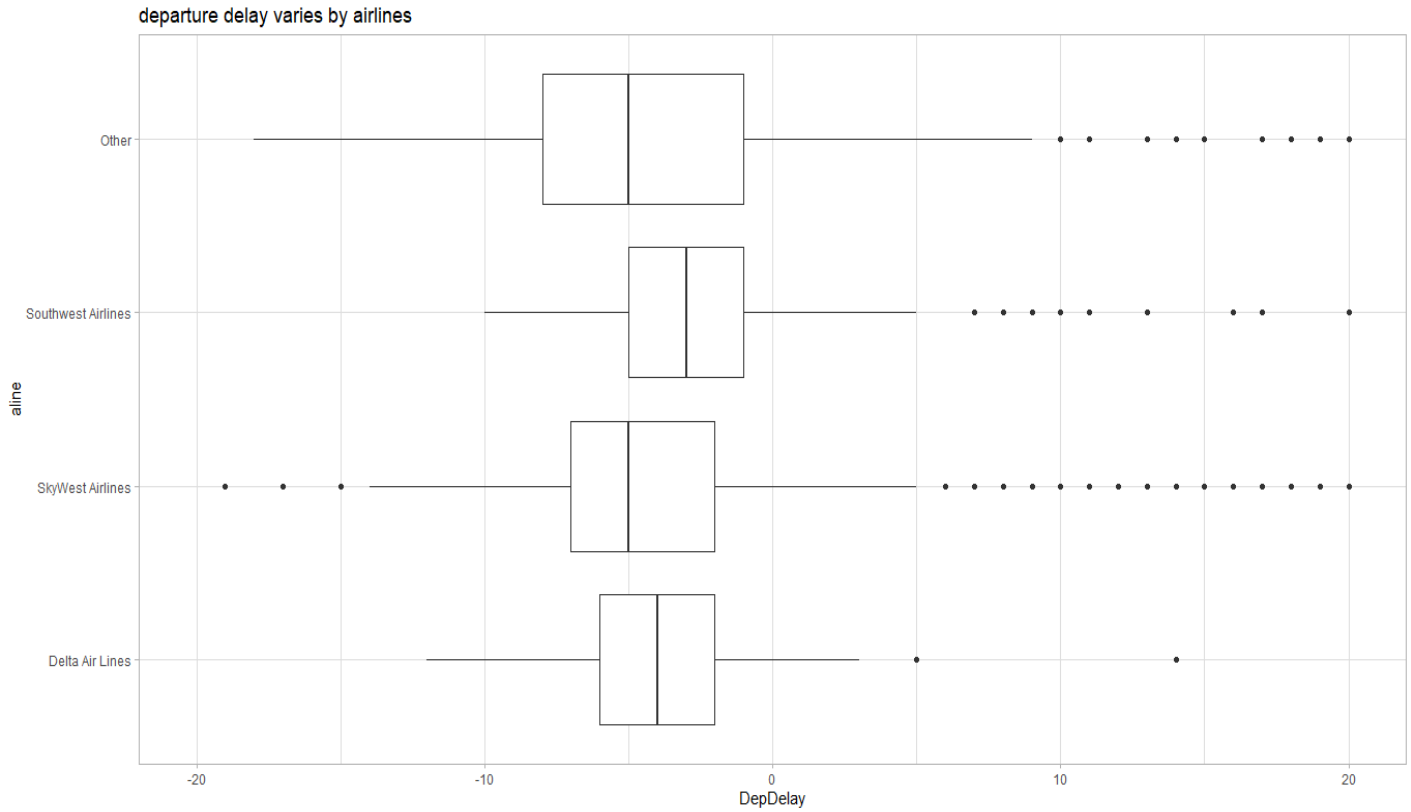
Exercise: Create a graphically excellent plot that shows the distribution of arrival delays.

Answer:



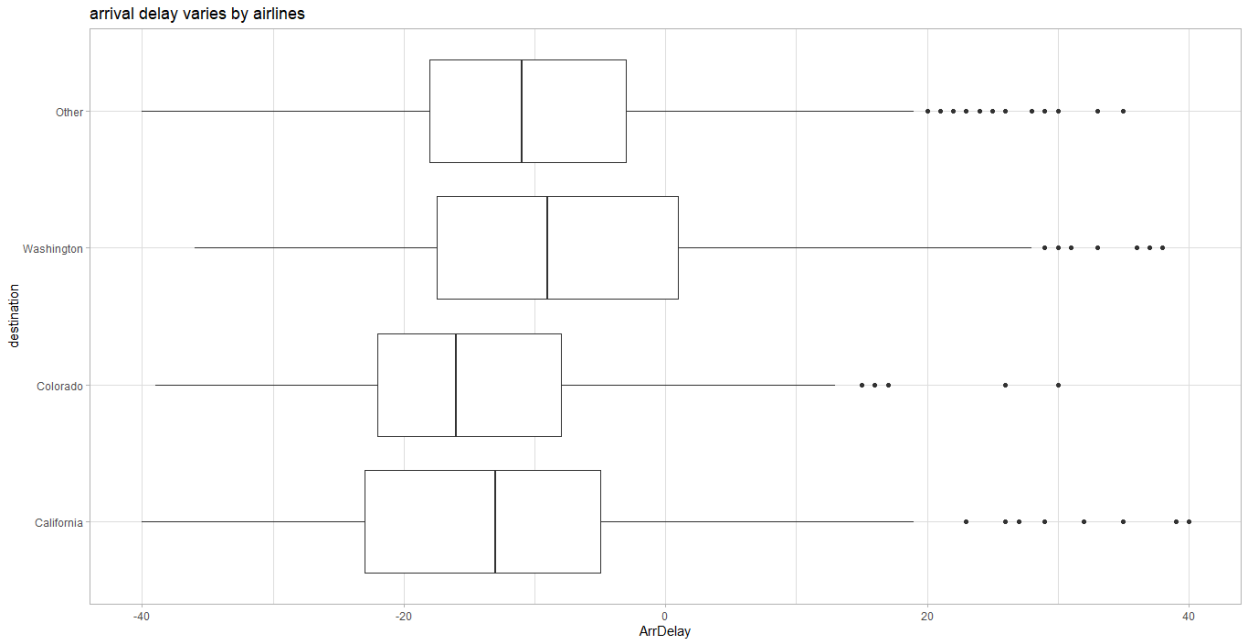
Exercise: Create a graphically excellent plot that shows how departure delay varies by airline. Include only the three most popular airlines and combine the rest as 'other'. Change your axis limits so you only display times between -20 and 20 minutes. *Hint: You may wish to use the `fct_lump` function from the `forcats` package.*

Answer:



Exercise: Create a graphically excellent plot that shows how arrival delay varies by destination. Include only the three most popular destinations and combine the rest as 'other'. Modify your axis limits so you only display times between -40 and 40 minutes. *Hint: You may wish to use the `fct_lump` function from the `forcats` package.*

Answer:



A3 Initial Interpretation (5%)

Exercise: Based on your graphical summary, does it appear that departure delay differs by airline?

Answer: Box plot of departure delay shows that median is below 0 which mean more that 50% flights departed early or on time. All boxes varies less and more consistent, where southwest and delta air lines median are in exact middle results more consistent predictions. However, There are outliers but SkyWest airline has high ratio of highest and lowest values of outliers whereas others has one end highest values of outliers.

Exercise: Based on your graphical summary, does it appear that arrival delay differs by destination?

Answer: Box plot of arrival delay shows that median is below 0 which mean more that 50% flights arrived early or on time. All boxes varies less, more consistent and make more dependable predictions. However, outliers are all having higher values can affect the predictions.

Exercise: Based on your graphical summary, does it appear that there is a linear relationship between departure delay and arrival delay?

Answer: The scatterplot shows a strong, positive, linear association between departure delays and arrival delays. The plot shows some potential outliers in the data .

Section B - Hypothesis Testing (30%)

In this section, you will perform hypothesis testing to determine whether the proportion of late flights is independent of airline. The level of significance here is $\alpha = 0.05$.

Section B1 - Setup (15%)

Exercise: Consider the two most popular airlines in your dataset. Remove any observations with missing departure delays. Fill in the table below with the counts of late flights and not late flights (early/on time).

Answer:

Airline	Late	Not late	Total
OO	182	735	917
WN	86	275	361
Total	268	1010	1278

Exercise: When testing for independence, we have learnt about both the χ^2 test and Fisher's Exact test. In this situation, which test is more appropriate?

Answer: Even though, Fisher's exact test gives the exact p-value and is more precise; it is used for small sample size. However, our sample size is quite huge and approximate p value received will not be exact but close enough to show relevance. And as result will be almost equal approximate, no need to put extra work on complex calculations as saving resources like time and hard work unless the retrieved result needs to be exact.

Exercise: What are the hypotheses for this test?

Answer: H0 (null-hypothesis): proportion of late flights is not depended on airlines.

H1: There is some dependence of late flights on airlines.

Exercise: Fill in the table below with the **expected** counts of late flights and not late flights (early/on time).

Answer:

Airline	Late	Not late	Total
OO	192.29734	724.7027	917.00004
WN	75.70266	285.2973	360.99996
Total	268	1010	1278

Section B2 - Hypothesis Test and Interpretation (15%)

Exercise: If you selected the χ^2 test, calculate the test statistic and the degrees of freedom.

Answer: Tests statistic = 2.470074, Degrees of Freedom = 1

Exercise: If you selected the χ^2 test, what is the p-value?

Answer: P-value: 0.1348

Exercise: What can you conclude based on this test?

Answer: The level of significance here is$\alpha = 0.05$. If p<α, we can should reject the null hypothesis. But, as p>α we accept the null hypothesis which states that late delay flights are not depended upon or are independent of popular airlines.

Section C - Linear Model (30%)

In this section, you will perform linear regression to examine the relationship between arrival delay and departure delay.

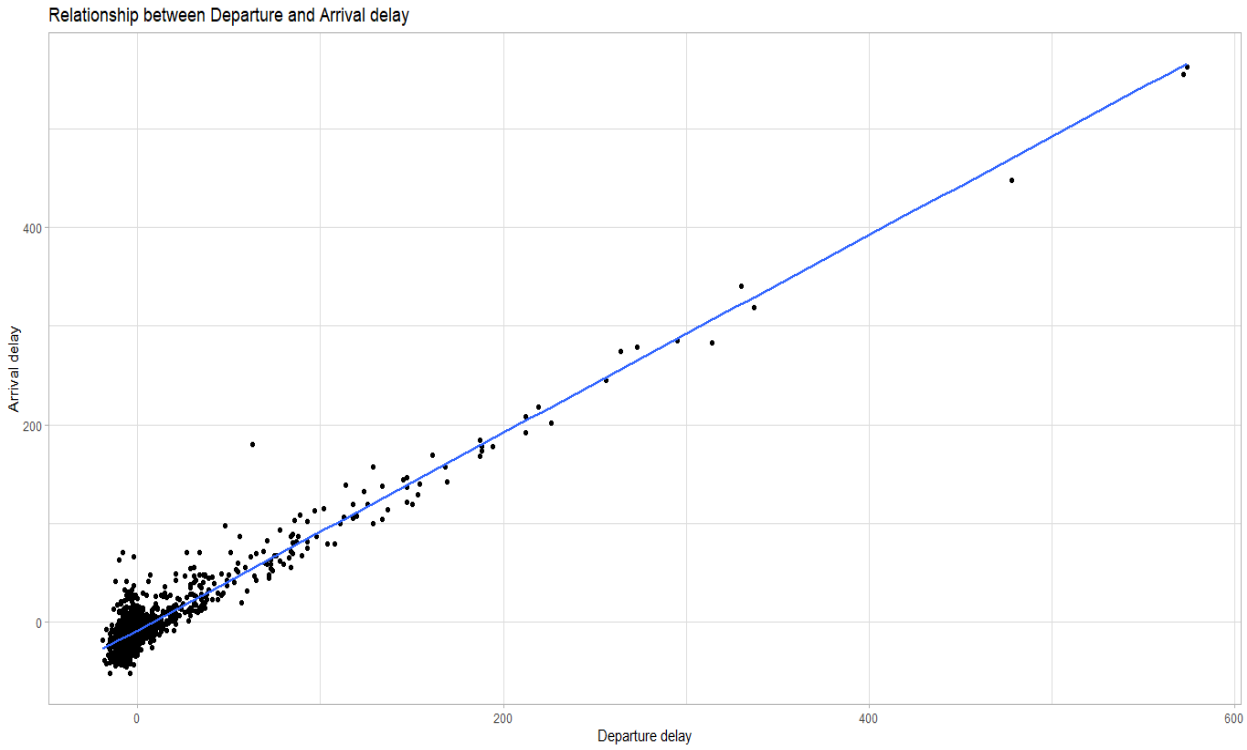
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where ε_i is Normally distributed with mean of 0 and variance of σ^2.

Section C1 - Create Linear Model (10%)

Exercise: Complete the following sentences so that they refer to the terms in the regression model above.

Answer: A linear model was specified to examine how the arrival delay, y, is related to the departure delay, x. The parameter b1(slope) describes the rate of change of the causality. The parameter b0 represents the arrival delay when x = 0 .



Exercise: Using R, fit a linear model to your chosen variables. Produce a captioned, well-formatted table below that includes a descriptive parameter name, estimate and 95% confidence interval.

Answer:

```
Call:
lm(formula = ArrDelay ~ DepDelay, data = lr_df)

Residuals:
    Min      1Q  Median      3Q     Max
-39.939  -8.935  -0.940   6.061 125.068

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.061040   0.329866  -24.44   <2e-16 ***
DepDelay     0.999884   0.008034  124.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.26 on 1650 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.9037,    Adjusted R-squared:  0.9037
F-statistic: 1.549e+04 on 1 and 1650 DF,  p-value: < 2.2e-16

> |
```

```
> get_regression_table(linmod)
# A tibble: 2 x 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept    -8.06     0.33      -24.4       0   -8.71     -7.41
2 DepDelay      1        0.008     124.        0    0.984     1.02
>
```

Exercise: Substitute your parameter estimates into the linear model below.

Answer: : $\hat{y}$ = -8.06 + (1) x + ε_i

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

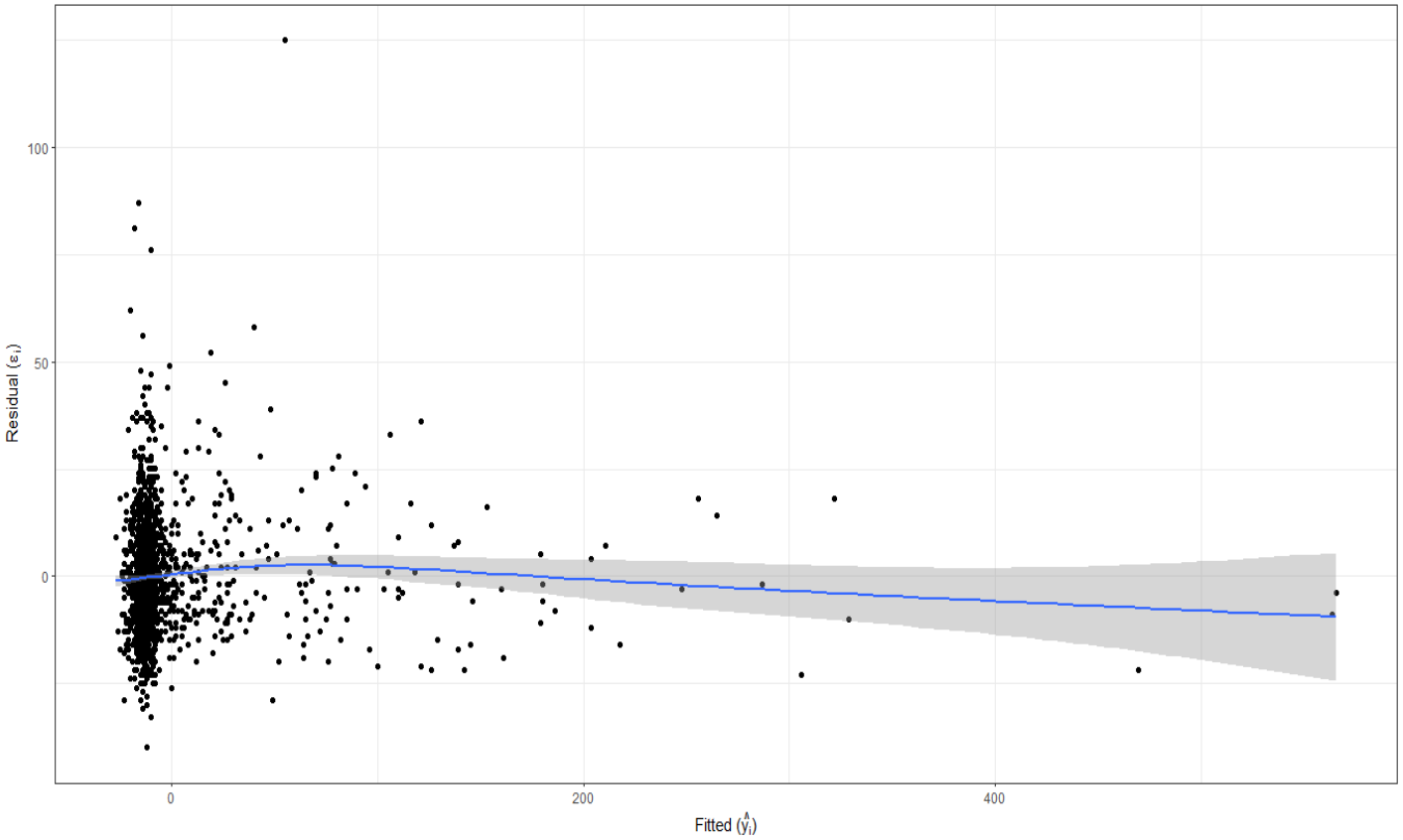
Exercise: How much variability in the observed data does your model explain?

Answer: The model explains 90.37% of the variability in the data.

Section C2 - Regression Assumptions (20%)

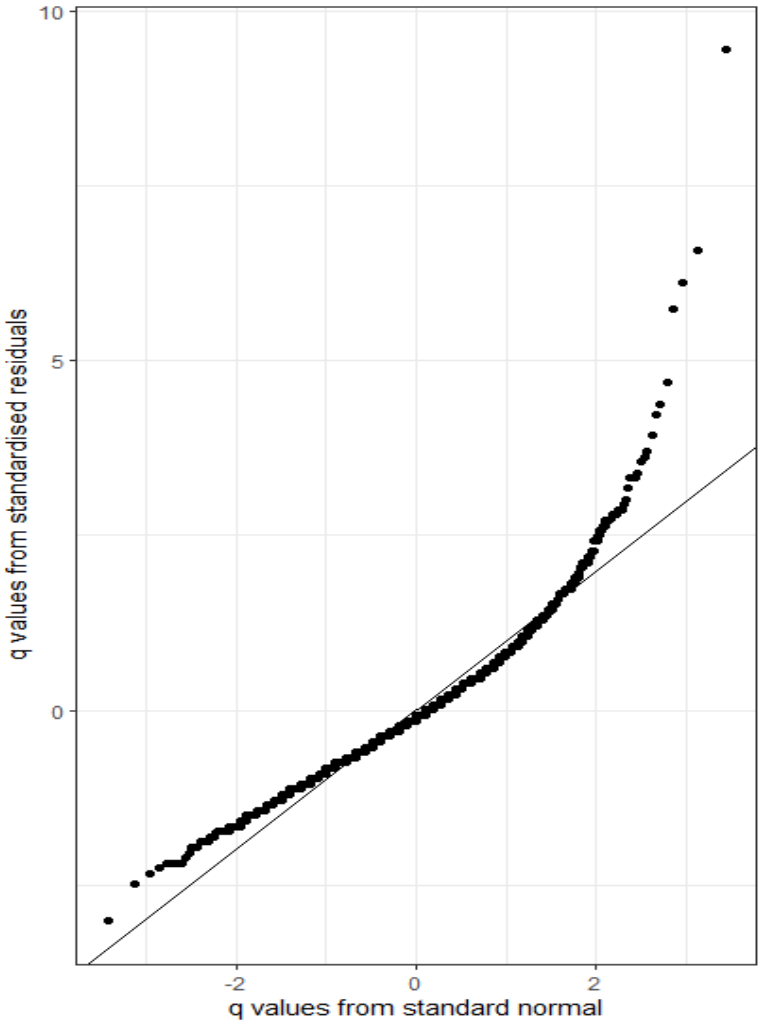
Exercise: Create a plot that shows how the residuals vary with the values fitted through your regression model.

Answer:



Exercise: Create a QQ plot that compares the standardised residuals to a standard normal distribution.

Answer:



Exercise: Based on these plots, does it appear that the residuals are normally distributed?

Answer: Residual show slight curve, however most of the residuals are close to line hence it does shows normal distribution.

Now, you will complete a Kolmogorov-Smirnov (KS) test to compare your standardised residuals to a standard Normal distribution.

Exercise: State the hypotheses for this test.

Answer: Ho: Standardized residuals follows standard normal distribution.

H1: Standardized residuals does not follow standard normal distribution.

Exercise: Plot the standard normal cumulative density function (CDF) and the empirical CDF of the standardised residuals.

Answer:

Exercise: Use the `ks.test` function in `R` to perform a KS goodness of fit test. Report the p-value.

Answer:

```
data:  IQ$IQ
D = 0.2331, p-value = 0.007149
alternative hypothesis: two-sided

> |
```

Exercise: Interpret the p-value you obtained from the KS test. What does this mean in terms of your hypotheses? Does the model for arrival delay as a function of departure delay satisfy the assumptions of linear regression?

Answer: p<α, hence, we reject the null hypothesis. Standard residuals does not follow normal distribution.

Section D - Interpretation (10%)

Write 1-3 sentences on each of the following, referring to the results of your analysis where necessary. If you rely on sources of information other than your analysis, ensure you cite them and that a bibliographic entry is provided below.

Exercise: In Section C, you presented a linear model to describe the relationship between departure delay and arrival delay. Explain what $beta_0$ and $beta_1$ mean in a real-world context. In other words, what does your model mean?

Answer: 1). The beta mean shows changes in means for E(y) that is predictor variable increase per unit when all variables are constant. Eg : $beta_1$ show change of mean for E(y) when all xi are hold constant. $beta_0$ does the same except when xi are not constant but zero. For example in real world, we relate dependent variable income depend upon independent variable education. In model, b1 quantifies effect of education on income, whereas, bo depicts as aconstant which at least depicts minimum income for person with no education.

Exercise: Your model considers all airlines simultaneously. Is this a limitation? What could be gained from considering airlines separately?

Answer: Model considers multiple variability of one variable. That's not a limitation. You can even use multiple regression. The limitation occurs only when data is not sufficient or correlation has concluded as causation. More relative predictions and identifying anomalies or outliers.

Exercise: If you found that the model for arrival delay as a function of departure delay does not satisfy the assumptions of linear regression, what should be done next?

Answer: First you can use logarithmic scale instead as they same up to linear scaling because smallest of changes in log results into percentage change. Another way is to add one more repressor which is nonlinear of one variable. At last you might need to introduce entirely independent variable that corrects the assumptions and non-linearity among variables.

Exercise: In real life, what other factors may influence the arrival delay experienced by a flight?

Answer: There are numerous reasons for flight to arrive late airtraffic control, security clearance, adverse weather conditions, late departure, strikes, emergency landing somewhere, etc.

Exercise: Assuming the linear model does hold, over what range would this model be valid? What are the limitations of this model?

Answer:

Refrences:

1) STAT 462, Applied Regression Analysis (2018), The Pennsylvania State University. https://online.stat.psu.edu/stat462/node/131/
2) David Weedmark (2018)The Advantages & Disadvantages of a Multiple Regression Model. https://sciencing.com/advantages-disadvantages-multiple-regression-model-12070171.html
3) Regression diagnostics: testing the assumptions of linear regression , Robert Nau, Duke, University. http://people.duke.edu/~rnau/testing.htm
4) Lecturer, Belinda (2020). MXN500 Statistical analysis: [Hypothesis Testing, Linear Regression]. QUT Blackboard.
5) 360 DataScience Course program, Online, Intro. To R programming and Statistics. https://365datascience.com/