Section A - Data and Visualisation

A1 Data Structure

1. A data dictionary of the variables in the data frame.

Variable	Description	Type	Class	
FlightDate	Scheduled date of the flight (dd/mm/yyyy)	Interval	Character	
Reporting_Airline	Airline undertaking the scheduled flight	Nominal	Character	
Origin	Airport code of the origin airport	Nominal	Character	
OriginStateName	Name of the state of the origin airport	Nominal	Character	
Dest	Airport code of the destination airport	Nominal	Character	
DestStateName	Name of the state of the destination airport	Nominal	Character	
DepDelay	Difference between scheduled and actual departure time (in minutes). (Negative numbers indicate early departures)	Interval	Numeric	
ArrDelay	Difference between scheduled and actual arrival time (in minutes). (Negative numbers indicate early arrival)	Interval	Numeric	
Airport Name	Name of the destination airport	Nominal	Character	
Airline name	Name of the airline	Nominal	Character	

2. Converting the data to more appropriate classes.

Variable	From	To
Reporting_Airline	Character	Factor
FlightDate	Character	Date

3. Summarising the number of observations, mean, median, and standard deviation of departure delays.

Number of observations	4173
Mean	2.245

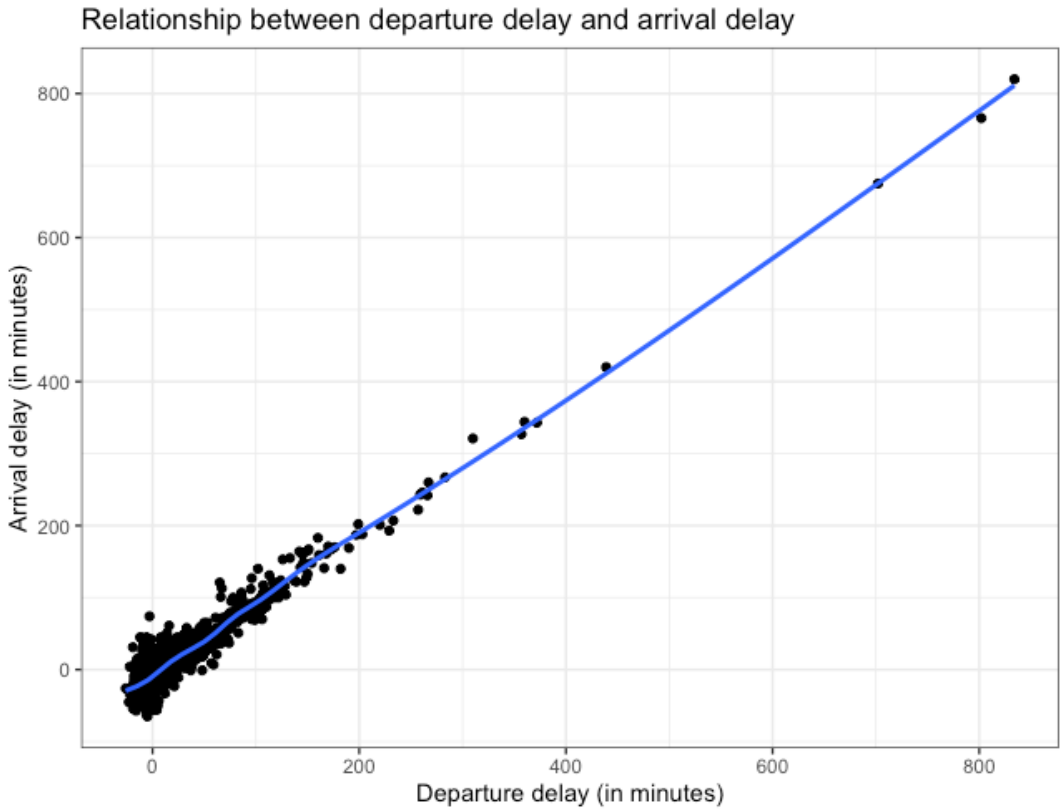
Median	-5.000
Standard deviation	35.78437

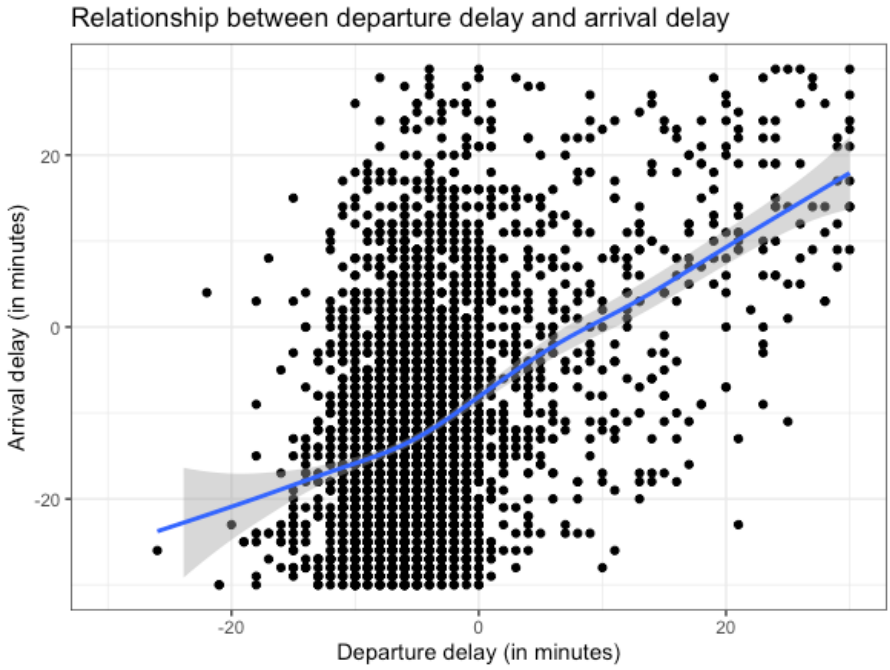
4. Reporting the data for each of the five most popular airlines.

5 most popular airlines	YX	WN	AA	9E	DL
Number of observations	889	849	434	294	286
Mean	-1.291383	1.152096	3.800000	5.672355	-1.300699
Median	-6	-3	-6	-5	-5
Standard deviation	25.58567	21.29995	59.41770	47.35268	16.30845

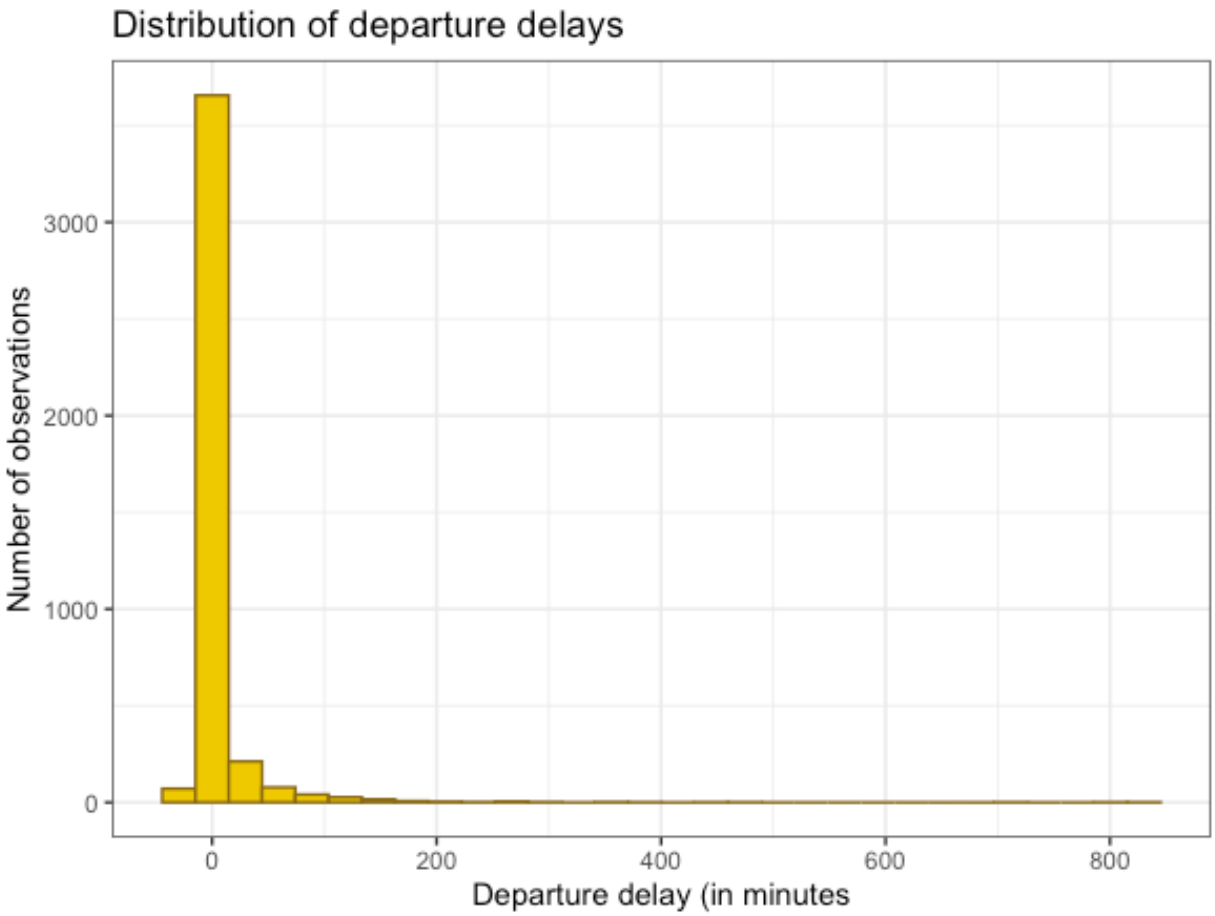
A2 Graphical Summaries

1. Plot that shows the relationship between departure delay and arrival delay.

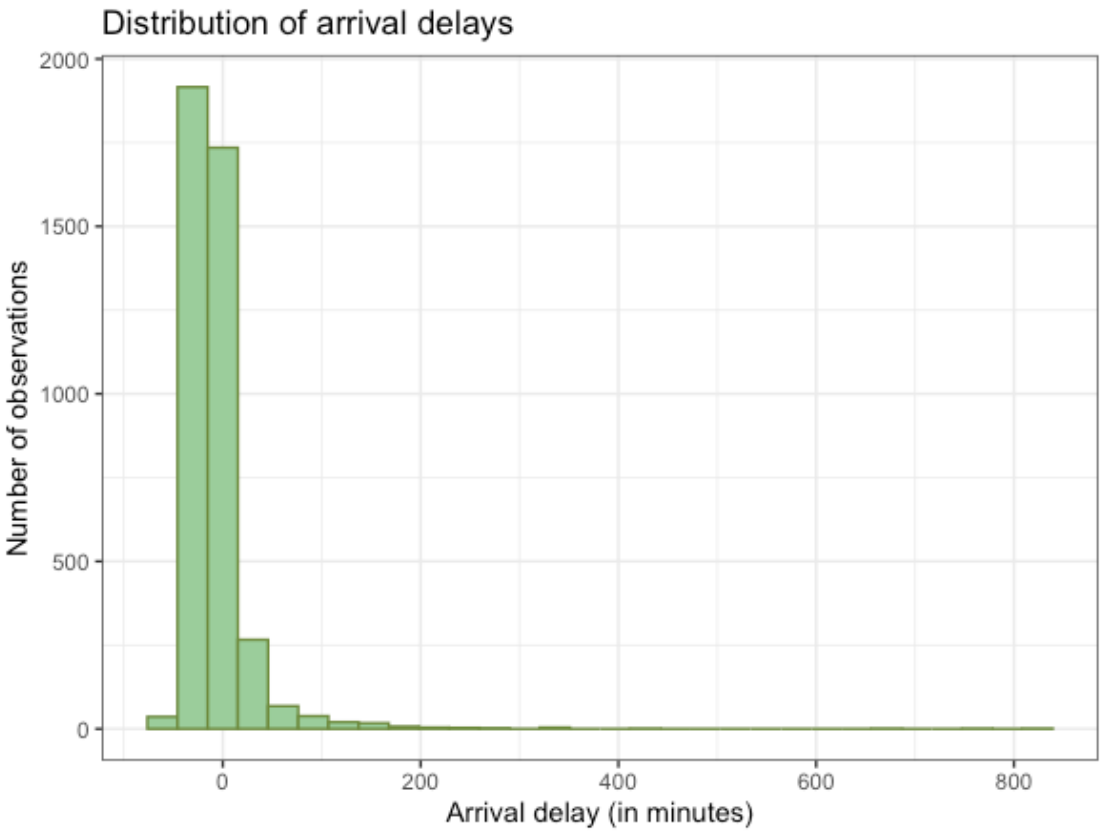

Relationship between departure delay and arrival delay
Arrival delay (in minutes)
800
600
400
200
0
0
200
400
600
800
Departure delay (in minutes)


Relationship between departure delay and arrival delay
Arrival delay (in minutes)
20
0
-20
-20
0
20
Departure delay (in minutes)

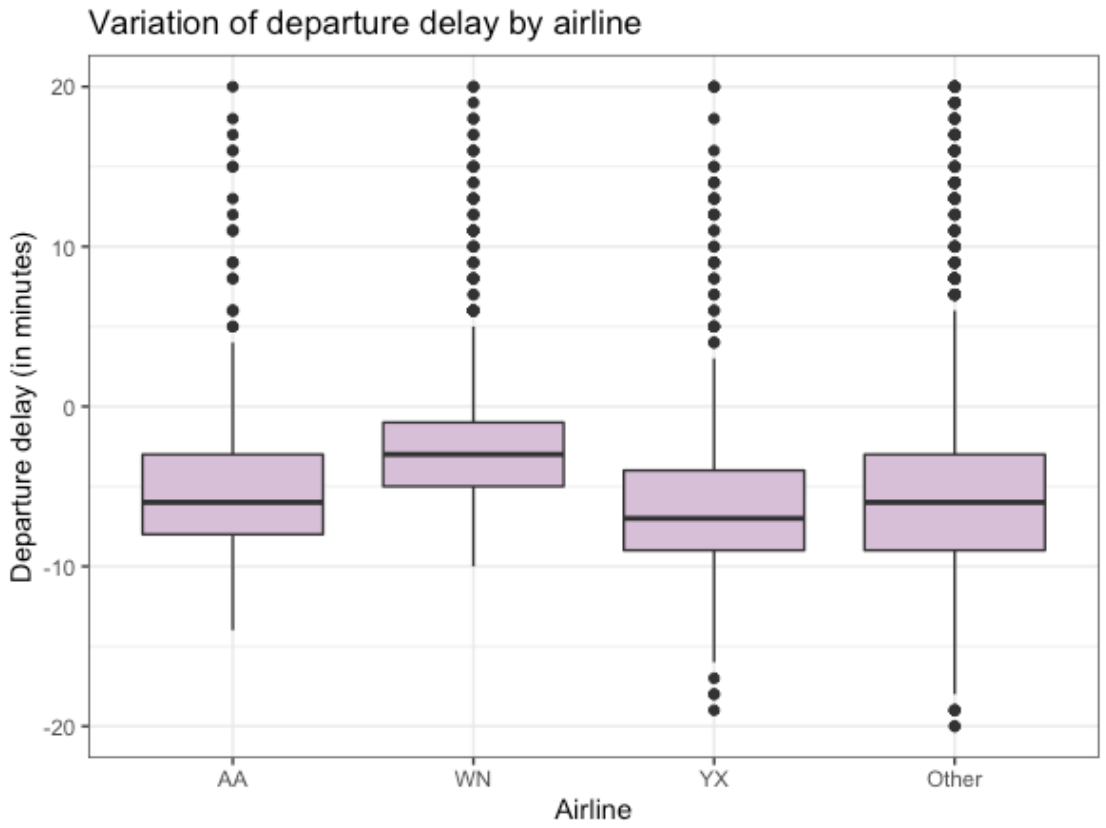
2. Plot that shows the distribution of departure delays.



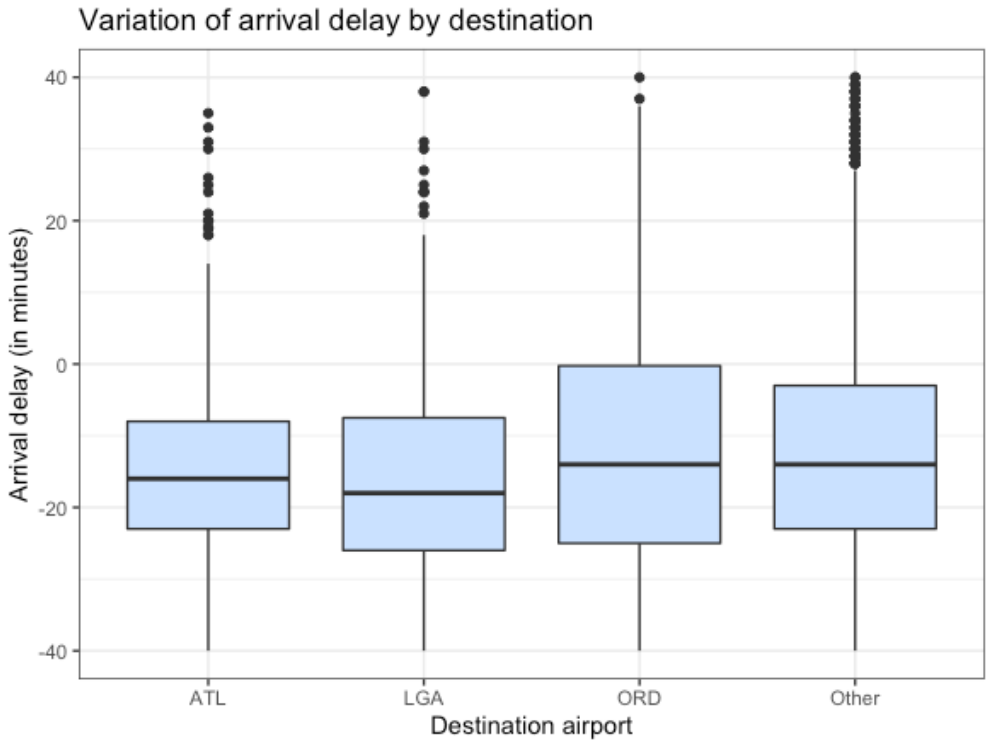
3. Plot that shows the distribution of arrival delays.



4. Plot that shows how departure delay varies by airline. Include only the three most popular airlines and combine the rest as 'other'.



5. Plot that shows how arrival delay varies by destination. (Including only the three most popular destinations and combining the rest as ‘other’.)



A3 Initial Interpretation

- The departure delay seems to differ by airline as the box-plot graph shows different mean for the top 3 airlines
- The arrival delay seems to differ by destination as the box-plot graph shows different mean for the top 3 destinations.
- As can be inferred from the graph above, there is a linear relationship between departure delay and arrival delay times. In order to get a better understanding of this relationship, we limit the scale between -30 and 30 minutes for both axes, where majority of the data is.
- In the graph above, the approximate y-intercept is at (0,-10) and the x-intercept at (10,0). This means that for a plane departing on time, it is likely to arrive at or before time.

Section B - Hypothesis Testing

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

1. Considering the two most popular airlines in your dataset. Removing any observations with missing departure delays. Filling in the table below with the counts of late flights and not late flights (early/on time).

Airline	Late	Not late	Total
WN	178	653	831
YX	110	771	881
Total	288	1424	1712

2. χ^2 test and Fisher's Exact test. In this situation, which test is more appropriate?

   χ^2 test. While the fisher's exact test is used to test the association between two categorical variables with small cell sizes, it is more prone to Type 1 error. Also, both tests give similar results for large datasets.

3. The hypotheses are:

   H_0: There is no association between the proportion of late flights and the type of airline.

   H_1: There is some association between the proportion of late flights and the type of airline.

4. The **expected** counts of late flights and not late flights (early/on time).

Airline	Late	Not late	Total
WN	139.7943	691.2056	830.9999
YX	148.2056	732.7934	880.9990
Total	287.9999	1423.9990	1711.9989

Section B2 - Hypothesis Test and Interpretation

Test statistic = 845.6785

Degrees of freedom = 3

p- value = 5.361571e-183

As p-value is less than $\alpha = 0.05$, we can conclude that we have enough evidence to reject hypothesis H_0 that there is no association between the proportion of late flights and the type of airline.

Section C - Linear Model

In this section, we 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

A linear model was specified to examine how the arrival delay, y_i is related to the departure delay, x_i . The parameter β_1 describes the rate of change of the arrival delay. The parameter β_0 represents the arrival delay when departure delay is zero.

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

term	estimate	conf.low	conf.high	p.value
(Intercept)	-9.54	-9.97	-9.11	0
DepDelay	0.995	0.982	1.01	0

- Substituting your parameter estimates into the linear model below.

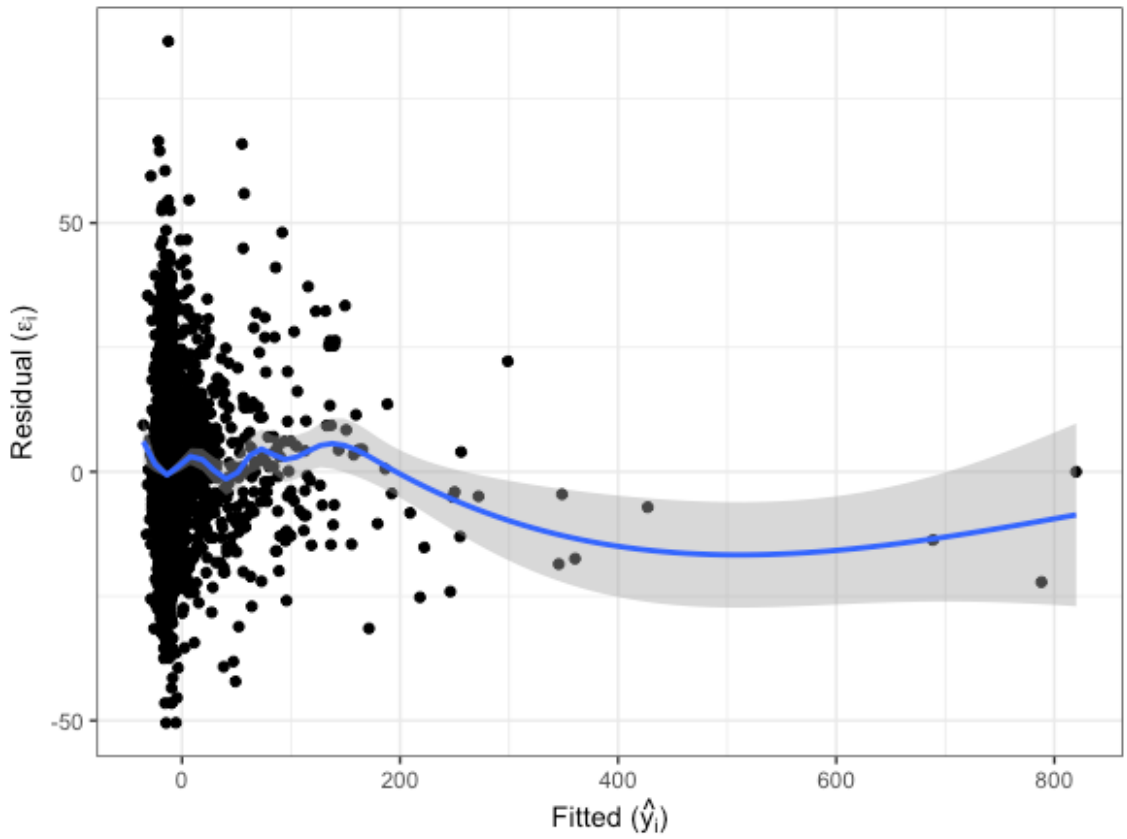
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

$$y_i = -9.54 + 0.995 x_i$$

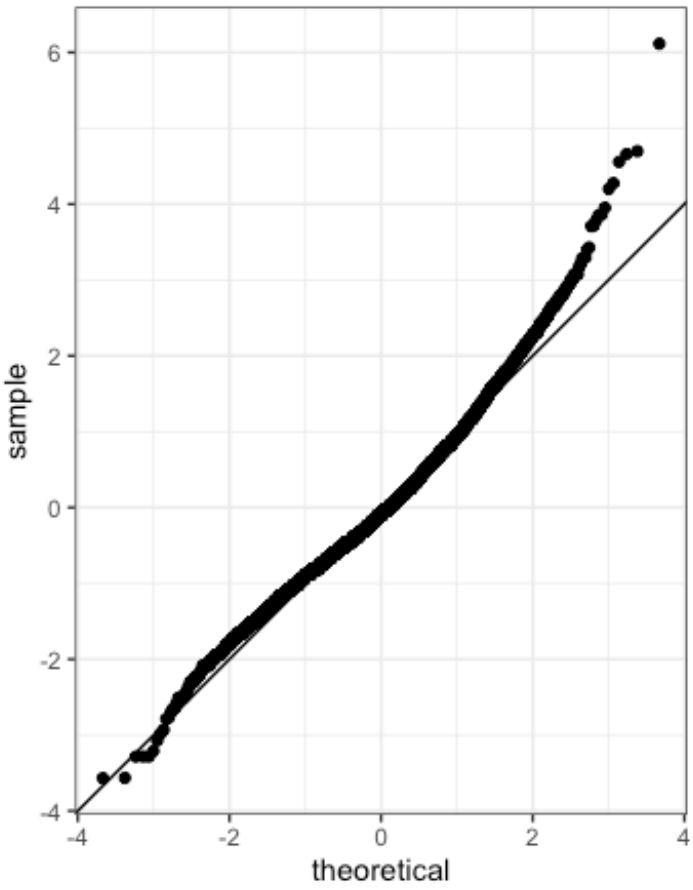
- The R-squared value of 0.8584 represents that the variability of the observed data of the departure delay model is 85.84%.

Section C2 - Regression Assumptions

1. A plot that shows how the residuals vary with the values fitted through your regression model.



2. A QQ plot that compares the standardised residuals to a standard normal distribution.



- As observed in the above plots, the residual points are closer to the line, excluding a few outliers to the right. This shows that the data is normally distributed.

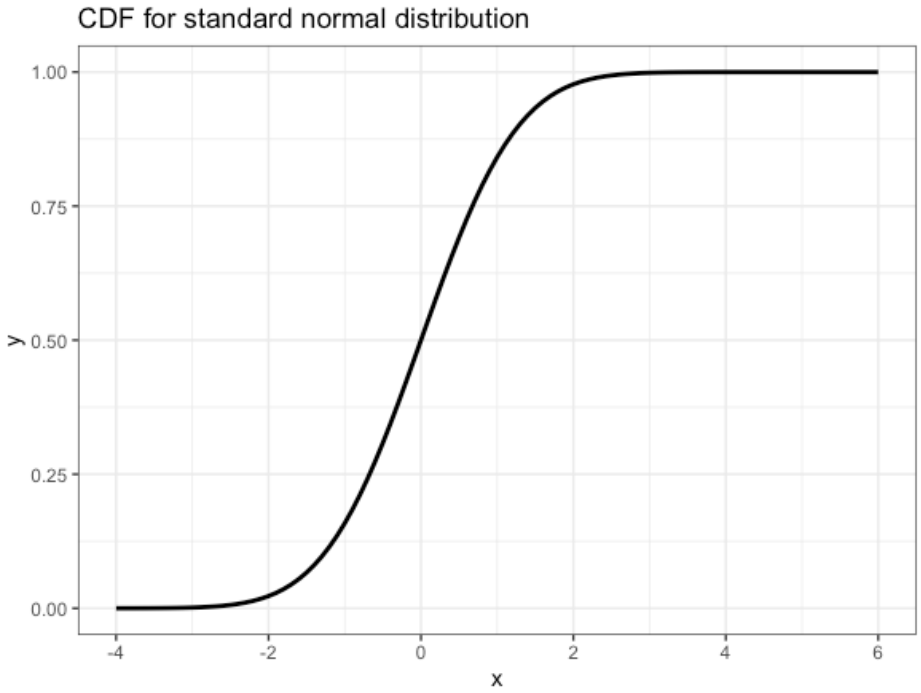
Kolmogorov-Smirnov (KS) test to compare the standardised residuals to a standard Normal distribution.

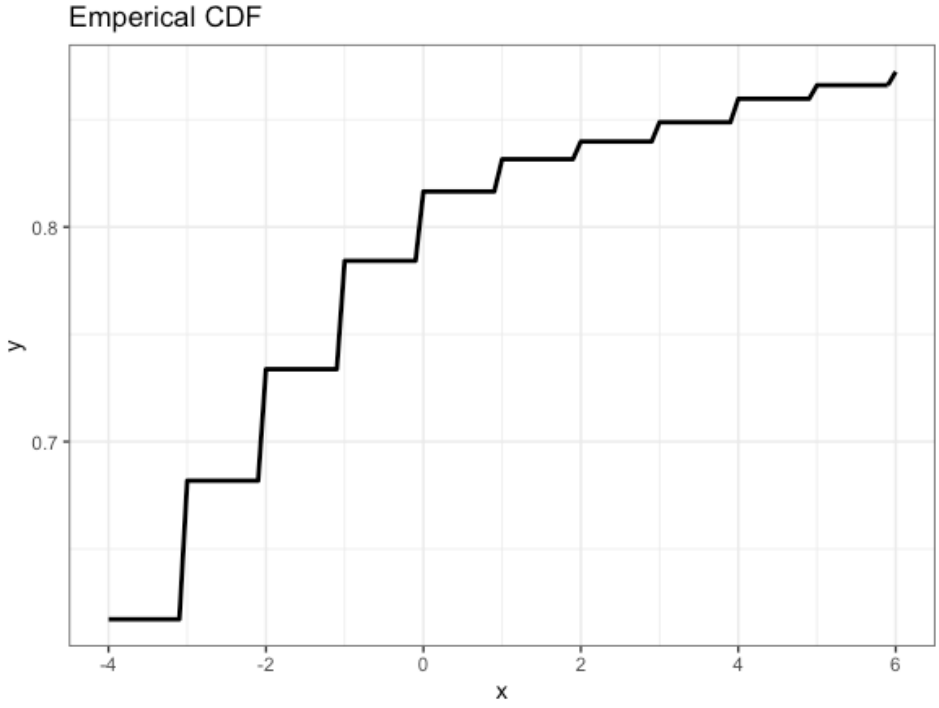
The hypotheses are:

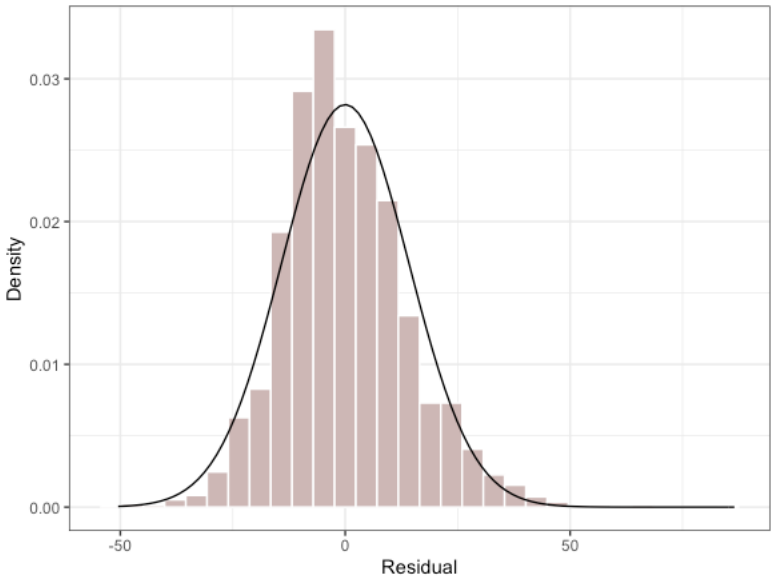
H_0: The standardized residuals follow a normal distribution.

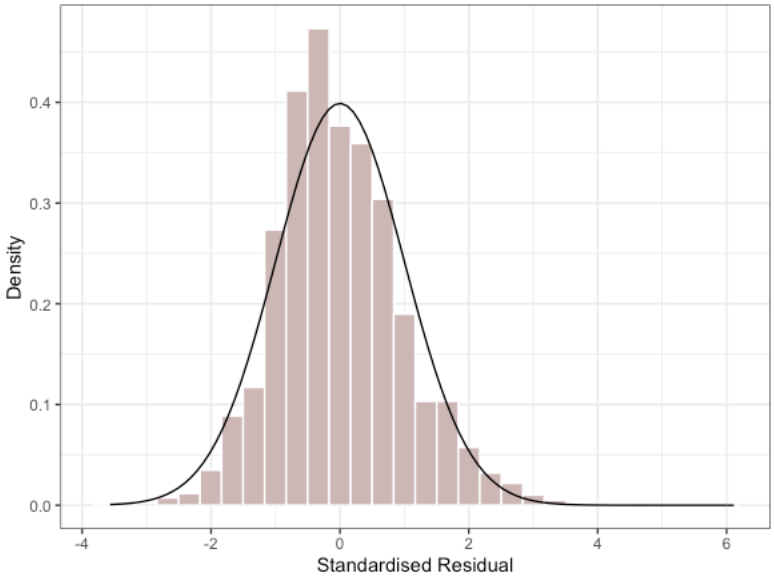
H_1: The standardized residuals do not follow a normal distribution

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


CDF for standard normal distribution
1.00
0.75
0.50
0.25
0.00
y
-4
-2
0
2
4
6
x


Emperical CDF
0.8
0.7
y
-4
-2
0
2
4
6
x





2. KS goodness of fit test. Report the p-value.

```
> p_cdf <- function(x){return(pnorm(q = x, mean = 100, sd = 15))}
> # KS goodness of fit test
> ks.test(x = dep.fort$DepDelay, y = "p_cdf")

        One-sample Kolmogorov-Smirnov test

data:  dep.fort$DepDelay
D = 0.96143, p-value < 2.2e-16
alternative hypothesis: two-sided
```

On performing KS goodness of fit test, we get p-value < 2.2e-16.

- So, p-value from KS test is < 2.2e-16. As $p < 0.05$ we can say that there is significant evidence against the hypothesis H_0 that standardized residuals follow normal distribution.