Report of Data Science Coursework

By: Marcus Ting En Hao

R

Question 1

Pre-set up

1.I need to load some libraries required in order to work with some features that otherwise won't which are DBI and dplyr

2. I set a working directory, which allows me to load and save data from and back using setwd()

3. I load the data of 2007 and 2008 into the dataframes from CSV using read.csv()

4.. Afterwards i created a database which allow me to extract information when i need, but if there already exist one, i will remove it before i create a new one. Using if(file.exists())file.remove() and dbconnect(RSQLite::SQLite(),)

5. I combine the data of 2007 and 2008 into 1 single dataframe using rbind(,)

6. I load the data of previous step into the database using dbwritetable()

Loading the answers

7. Each of the following answers uses dbGetQuery to get data from the database, and using the function WHERE to include data only if Cancelled and Diverted are 0. This is because there are no delays where flights are cancelled or diverted, which will spoil the answer. Each answers uses the average delay to find out the respective questions. I use the sum of Depdelay, Arrdelay, CarrierDelay, Weatherdelay,NASdelay,Securitydelay and late aircraft delay for the total. Each answers are also group by their respective questions.

7.To get the best time to fly with minimal delay, i link the departure times of 2007 and 2008 to the average delay of that specific flight. This will give an answer of 551hrs

8. To get the best day of the week to fly with minimal delay,i link every day of the week of 1-7 from 2007 and 2008 to the average delay of that specific flight using the AVG() function. This will give an answer of 6, which is Saturday

9. To get the best year to fly with minimal delay, i link the years to the average delay of those years. This will give an answer of 2007

Question 2

Pre-set up

1.I need to load some libraries required in order to work with some features that otherwise won't which are DBI and dplyr

2. I set a working directory, which allows me to load and save data from and back using setwd()

3. I load the data of 2007,2008 and plane-data into the dataframes from CSV using read.csv()

4.. Afterwards i created a database which allow me to extract information when i need, but if there already exist one, i will remove it before i create a new one. Using if(file.exists())file.remove() and dbconnect(RSQLite::SQLite(),)

5. I combine the data of 2007 and 2008 into 1 single dataframe using rbind(,)

6. I load the data of previous steps and plane-data into the database using dbwritetable()

Loading the answer
7. I achieve this by geting the data of the average delay of the newest and oldest model by using the dbGetQuery() with AVG() function. I use the sum of Depdelay, Arrdelay, CarrierDelay, Weatherdelay,NASdelay,Securitydelay and late aircraft delay for the total delay. Using the WHERE function to include only Cancelled and diverted flights because there is no delay if flights are cancelled or diverted.

8. Once obtaining the data from dbGetQuery(), I put the model and average delay variables into buckets using the paste()

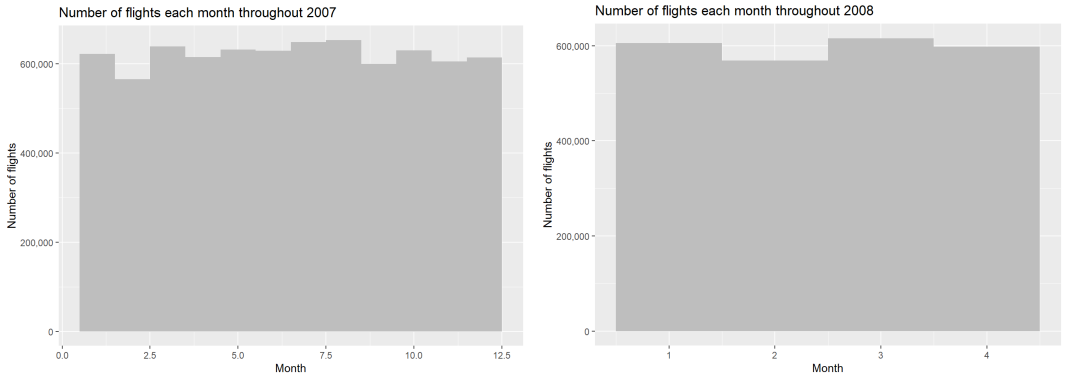
9. I converted the average delay buckets into numeric using the as.numeric()

10. I set the condition that if the older plane model has more average delay than the newer plane model, the answer will be TRUE, if not FALSE, using the if else function.

11. I print the answer would which states that the older model, which is the DC-7BF, has an average of 71.18 while the newer model, which is the 777-232LR, has an average of 108.75

Question 3
Pre-set up
1.I need to load some libraries required in order to work with some features that otherwise won't which are DBI,dplyr and ggplot2
2. I set a working directory, which allows me to load and save data from and back using setwd()
3. I load the data of 2007 and 2008 into the dataframes from CSV using read.csv()
4.. Afterwards i created a database which allow me to extract information when i need, but if there already exist one, i will remove it before i create a new one. Using if(file.exists())file.remove() and dbconnect(RSQLite::SQLite(),)
6. I load the data of previous steps and plane-data into the database using dbwritetable()

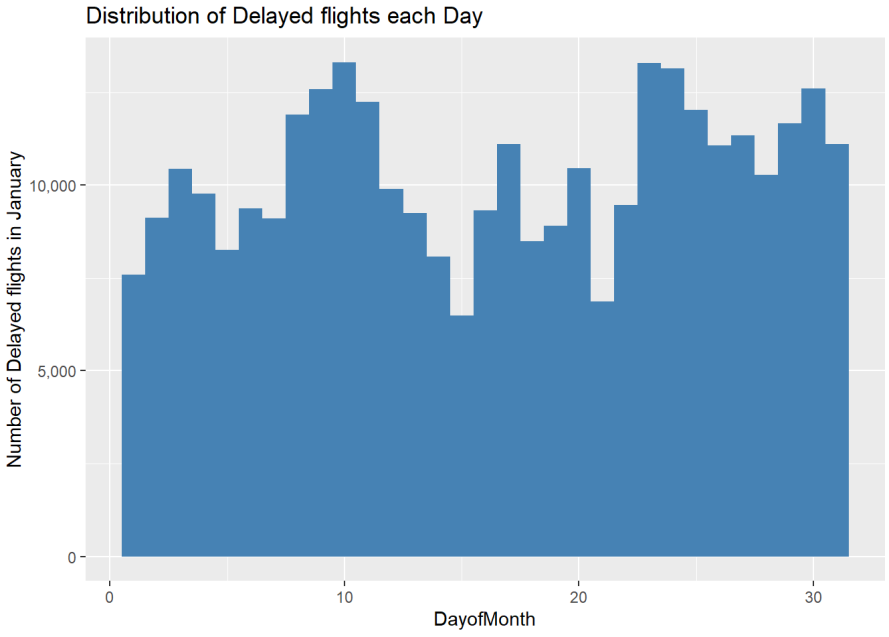
Loading the answer
7. I used ggplot function to create a histogram which describes the number of flights each month throughout 2007 and 2008.



This shows how the number of flights change throughout 2007 and 2008.

Question 4
Pre-set up
1.I need to load some libraries required in order to work with some features that otherwise won't which are DBI,dplyr and ggplot2
2. I set a working directory, which allows me to load and save data from and back using setwd()
3. I load the data of 2007 into the dataframes from CSV using read.csv()
4.. Afterwards i created a database which allow me to extract information when i need, but if there already exist one, i will remove it before i create a new one. Using if(file.exists())file.remove() and dbconnect(RSQLite::SQLite(),)
5. I load the data of previous step into the database using dbwritetable()
6.I create a new column which has the sum of all delays,which are Depdelay, Arrdelay, CarrierDelay, Weatherdelay,NASdelay,Securitydelay and lateaircraftdelay using the with(,paste0())
7.I create a new column where if the total delay is more than 0, a new variable 1 is created, if total delay is less or equal 0, a new variable 0 is created. Using the [df$]
8.I change the data into a single month using subset()
9.I updated the data to remove all rows which have total delay less than or equal to 0 using subet()

Loading the answer
10. Using ggplot2 to create a histogram which shows the Delayed flights each day for the month of Janurary



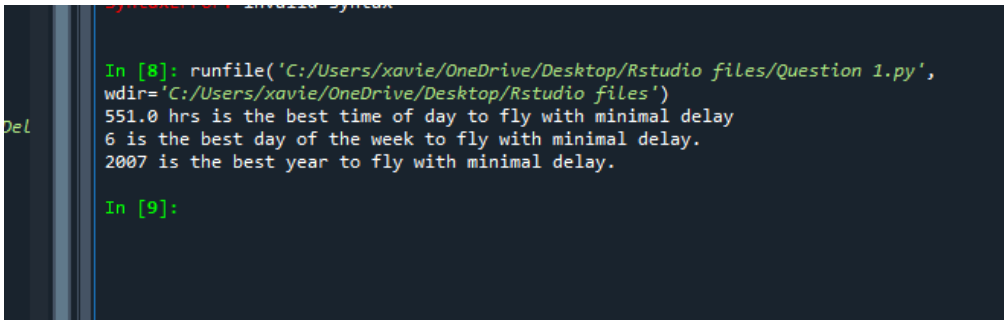
This graph shows that there is a higher chance that the next day have more delayed flights if the previous day has more delays than 7500 flights. Which proves that airport delays are cascading.

Question 5
Pre-set up
1.I need to load some libraries required in order to work with some features that otherwise won't which are mlr3,mlr3learners,mlr3pipelines and mlr3tuning
2. I set a working directory, which allows me to load and save data from and back using setwd()
3. I load the data of 2007 into the dataframes from CSV using read.csv()
6.I create a new column which has the sum of all delays,which are Depdelay, Arrdelay, CarrierDelay, Weatherdelay,NASdelay,Securitydelay and lateaircraftdelay using the with(,paste0())
7.I create a new column where if the total delay is more than 0, a new variable 1 is created, if total delay is less or equal 0, a new variable 0 is created. Using the [df$]
8.I change the data into Jan and Feb only using subset()
9.I update the data to remove rows that flights are diverted or cancelled using subset(). As logic states that there are no delays if flights are cancelled or diverted
10.I update the dataframes to remove columns irrelevant to the questions.Due to limited memory to run the machine learning and time taken to load.
11. Changed the column required for the question into factors using the factor()
12. Create the training and testing sets, where the training set is 50% of total data and the training sets are the difference.
13. Create the task using the TRUE FALSE as the target to determine if there is delay or no delay. Using the columns of origin, day of month and uniquecarrier as learning points. Using taskclassif$new
14. Create conversation if the data does not accept factors using po()
15. Creating hyperparameters using tnr() and trm()

Loading the answer
16. By using logistic regression we get the answer of 0.4015593
This means the the machine learning model has a 60% accuracy using logistic regression

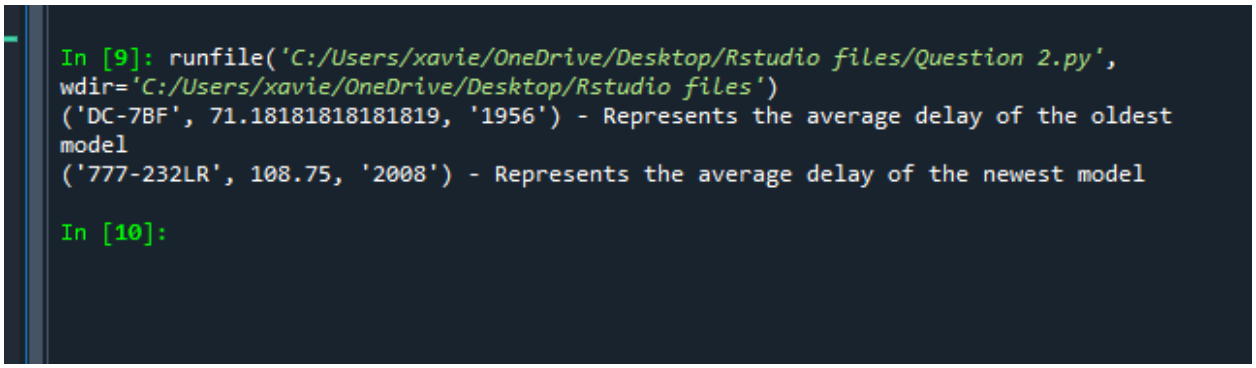
Python
Question 1
Pre-set up
1.I need to load some libraries required in order to work with some features that otherwise won't which are sqlite3, os and panda
2.Set working directory using the os.chdir()
3.If database already exist, remove it, else pass. Using the **try except** and **pass** function
4.connect to a database using sqlite3.connect().commit()
5.load the data of 2007 and 2008 using pd.read_csv()
6.combining 2 dataframes of 2007 and 2008 using .merge()
7.loading the data into the database using .to_sq1()

Loading the answer
8. Using connect.cursor.execute() as average while using the sum of Depdelay, Arrdelay, CarrierDelay, Weatherdelay,NASdelay,Securitydelay and late aircraft delay for the total and using the AVG(). GROUP by their respective question and order by average delay.



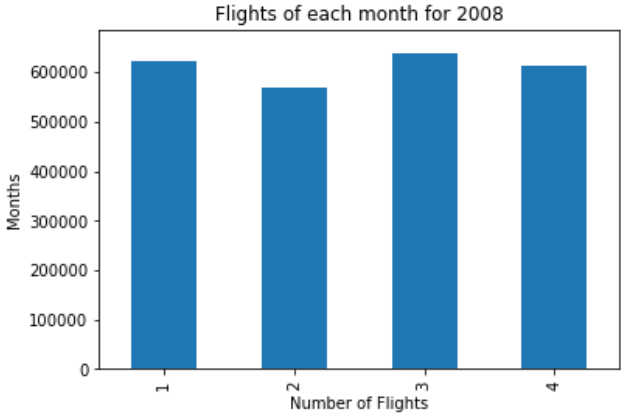
Question 2
Pre-set up
1.I need to load some libraries required in order to work with some features that otherwise won't which are sqlite3, os and panda
2.Set working directory using the os.chdir()
3.If database already exist, remove it, else pass. Using the **try except** and **pass** function
4.connect to a database using sqlite3.connect().commit()
5.load the data of 2007, 2008 and plane-data using pd.read_csv()
6.combining 2 dataframes of 2007 and 2008 using .merge()
7.loading the data into the database using .to_sq1()

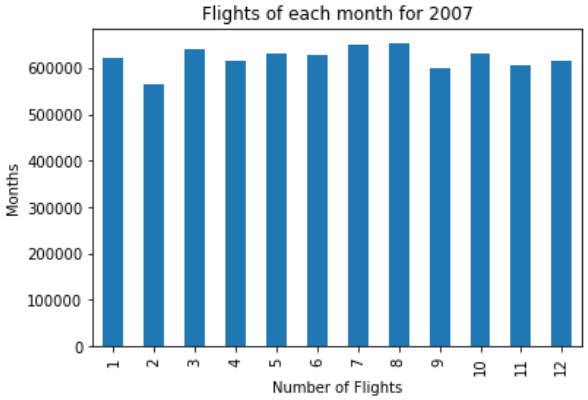
Loading the answer
8.
For the oldest model, i select column model as a link,using the sum of Depdelay, Arrdelay, CarrierDelay, Weatherdelay,NASdelay,Securitydelay and late aircraft delay for the total and using the AVG(). GROUP by model and order by the year. Linking the data USING the tailnum. Using WHERE to include only years more than 1900 and no cancelled or diverted flights



Question 3
Pre-set up
1.I need to load some libraries required in order to work with some features that otherwise won't which are matplotlib.pyplot, os and panda
2.Set working directory using the os.chdir()
3.load the data of 2007 and 2008 using pd.read_csv()

Loading answer
4.Using size().plot.bar() while grouping by Month, it displays the histograms of the number of flights from 2007 to 2009



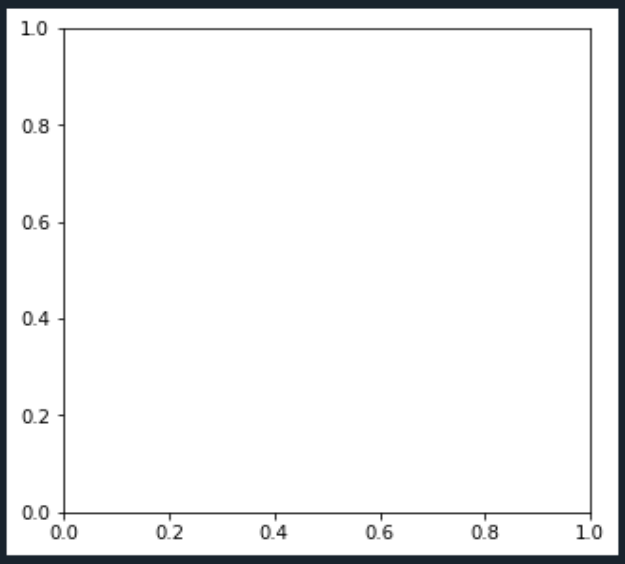


This show how the number of passengers change from the flights throughout the year 2007 and 2008

Question 4
Pre-set up
1.I need to load some libraries required in order to work with some features that otherwise won't which are numpy, os, panda, networkx and matplolib.pyplot
2.Set working directory using the os.chdir()
3.load the data of 2007 using pd.read_csv()
4.Creating a new column of TotalDelay using the sum of Depdelay, Arrdelay, CarrierDelay, Weatherdelay,NASdelay,Securitydelay and late aircraft delay.
5. Creating a new column with having more than 0 total delay, create a 1 vavirable if not create a 0 variable. Using the np.where()
6.Remove the columns which are not needed in the question using .drop()
7. Create a chart with Orgin and Dest as rows while the weight row to be TF
Loading the answer
It couldnt load as there was an error

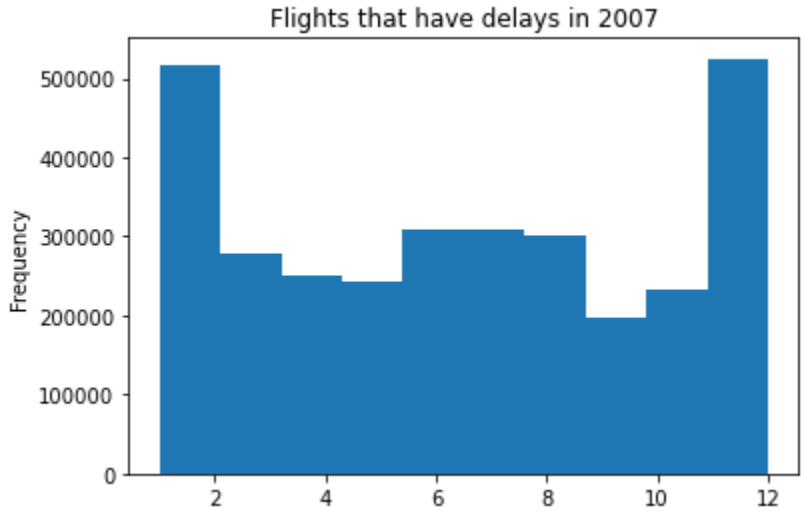
```
    if cf._axstack() is None:

TypeError: '_AxesStack' object is not callable
```




```
In [2]:
```

So instead i use a histogram
Where i drop all rows with false in the true false column using .drop()
Create a histogram with .plot



Shows that the airports are link and delays are cascading

Question 5

Pre-setup

1.I need to load some libraries required in order to work with some features that otherwise won't which are numpy, os, panda, SKLEARN and matplolib.pyplot.

2.Set working directory using the os.chdir()

3.load the data of 2007 and 2008 using pd.read_csv()

4.combining 2 dataframes of 2007 and 2008 using .merge()

5.Creating a new column of TotalDelay using the sum of Depdelay, Arrdelay, CarrierDelay, Weatherdelay,NASdelay,Securitydelay and late aircraft delay.

6.Creating a new column with having more than 0 total delay, create a 1 vavirable if not create a 0 variable. Using the np.where()

7. Features are uniquecarrier, origin dest, dayofmonth and TF

8. Turning dayofmonth into a numerical feature using numerical_transformer

9. Turning uniquecarrier, origin, tf and dest into categorical using categorical_transformer

10. Set the training and testing

Answer could not load as there is error