**ICT583 Data Science Applications**

**Murdoch University**

**Mid-term exercise - 30%**

**Instructions**

1. This exercise must be done **individually** by each student.
2. Write your answers in a report format. Clearly indicate each question/sub-question number, and give your code followed by the snapshot of your results and analysis if necessary.
3. You will submit .R files along with the report, so we can run it for check. Make sure your code exactly matches the provided answers. For example, if there are three separate data frames your code should produce the same three separate data frames.
4. Code should be easy to read and understand. Only include code and comments necessary for the exercise. Comments should clearly indicate each question/sub-question number.

*Do not rush into finding answers without a good understanding of the dataset. You should first ask yourself some questions like, what is this data frame about? What is the meaning of each dimension? What is the data type? Are there any dimensions describing similar things? Are there any missing values,etc.*

*All the data manipulation and visualization must be done using R.*

**Questions**

**Part One - small questions**

1. Use the **nycflights13** package and the **flights** data frame to answer the following questions:

1.1 What month had the highest proportion of canceled flights (the arr\_delay variable is NA)? What month had the lowest? Plot the proportion of canceled flights each month and interpret any seasonal patterns. (8 points)

1.2 What plane (specified by the tailnum variable) traveled the most times from New York City airports (JFK, LGA or EWR) in 2013? Plot the number of trips *per week*. (8 points)

1. Use the **Lahman** package and the **Teams** data frame to answer the following questions:
	1. Define two new variables in the Teams data frame: batting average (BA) and slugging percentage (SLG). Batting average is the ratio of hits (H) to at-bats (AB), and slugging percentage is total bases divided by at-bats. To compute total bases, you get 1 for a single, 2 for a double, 3 for a triple, and 4 for a home run. (8 points)

2.2 Display the top 15 teams ranked in terms of slugging percentage in MLB history. Repeat this using teams since 1969. (8 points)

2.3 Create a factor called election that divides the yearID into four-year blocks that correspond to U.S. presidential terms (from 1788 to 2017). During which term have the most home runs been hit? (Hint: seq function) (5 points)

1. Using the **storms** data frame from the **nasaweather** package:

Create a scatterplot between wind and pressure, with color being used to distinguish the type of storm. You might notice there are lots of overlapping data points in the scatterplot due to a comparatively large sample size, How would you improve your visualization? (5 points)

1. Using the **whately\_2015** data frame from the **macleish** package:

Create a data graphic that displays the average temperature over each 10-minute interval (temperature) as a function of time (when). Show both connected line and fitted line (8 points)

1. Suppose you are rolling two fair dies with success defined as getting a total value 4. If you roll two dies independently for eight times: (10 points)

3.1 What is the probability of observing exactly five successes (five total value 4s) in total? (calculated by hand)

3.2 Use R to confirm the result of Pr(X=5) for the die-roll example.

3.3 Plot the corresponding full probability mass function for X for this die-rolling example.

**Part Two - small projects**

COVID-19 Pandemic

The COVID-19 outbreak was first identified in December 2019 in Wuhan, China. The WHO declared the outbreak a Public Health Emergency of International Concern on 30 January 2020 and a pandemic on 11 March (Wikipedia). Organizations worldwide have been collecting data so that the government can monitor and learn from this pandemic. You will use the dataset ‘time\_series\_covid\_19\_confirmed.csv’ from LMS to explore the COVID-19 data. (20 points)

Note: This data set details can be found via https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset#time\_series\_covid\_19\_confirmed.csv;

Your data analysis should include but not limited to the answers to the following questions:

1. Create a clear bar chart that displays the latest number of COVID-19 cases of the top 10 countries. Consider how to improve the quality and aesthetics of your visualization.

2. Visualize the confirmed cases worldwide from January to March.

3. Visualize the confirmed cases of COVID-19 in China and the rest of the world from January to March. Can you relate the main changes observed from the plot with the landmark events such as WHO declared a pandemic?

4. Add a smooth trend line using linear regression to measure how fast the number of cases is growing in China after 15 February 2020. How does the rest of the world compare to linear growth?

5. Raise at least one question from your own regarding the COVID-19 pandemic and find answers using the given dataset.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Exploration of the bitcoin cryptocurrency market

After Bitcoin was launched in 2009, hundreds of similar projects based on the blockchain technology have emerged. Currently, Bitcoin is the world’s largest cryptocurrency by market capitalization. Meanwhile, the cryptocurrency market is exceptionally volatile and can make you lose all the money easily. You will use the dataset ‘cryptocurrency\_market\_2017.csv’ from LMS to explore the bitcoin cryptocurrency market.

Your data analysis should include but not limited to the answers to the following questions: (20 points)

1. Are there any cryptocurrencies listed in this dataset that have no known market capitalization (market\_cap\_usd)? If yes, they can be removed from the dataset.

2. Bitcoin has the largest market capitalization. Let’s compare Bitcoin with the rest of the cryptocurrencies. You can visualize the percentage of market capitalization for the top 10 coins as a barplot. Consider how to improve the plot to make it easier to read and convey more information.

3. Let’s explore the volatility of the cryptocurrencies market. You can select and plot the top 10 (sort by percent\_change\_24h in ascending order) coins’ 1 hour (percent\_change\_1h), 24 hours (percent\_change\_24h) and 7 days percentage change (percent\_change\_7d).

4. Design the bar plots to best display the daily and weekly biggest gainers and the biggest losers in market capitalization.

5. In general, cryptocurrencies with smaller capitalization are less stable projects, and therefore even riskier investments than the bigger ones. Following the Investopedia's large capitalization definitions (https://www.investopedia.com/terms/b/bitcoin.asp), let’s find out the coins with large capitalization.

6. Raise at least one question from your own regarding the bitcoin cryptocurrency market and find answers using the given dataset.