UNIVERSITY OF MALAYA

EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2017/2018 : SEMESTER I

WQD7001 : Principles of Data Science

Jan 2018 Time : 2 hours

INSTRUCTIONS TO CANDIDATES :

Answer **ALL** questions (50 marks).

ANSWER SCHEME

(This question paper consists of 4 questions on 4 printed pages)

1. Exploratory analysis is largely concerned with summarizing and visualizing data before performing formal modelling.
2. Indicate **FOUR** purpose of exploratory data analysis.

[4 marks]

**ANSWER:**

1. *Check for missing data and other mistakes.*
2. *Gain maximum insight into the data set and its underlying structure.*
3. *Uncover a parsimonious model, one which explains the data with a minimum number of predictor variables.*
4. *Check assumptions associated with any model fitting or hypothesis test.*
5. *Create a list of outliers or other anomalies.*
6. *Find parameter estimates and their associated confidence intervals or margins of error.*
7. *Identify the most influential variables.*
8. The snapshot of the dataset “mtcars” is shown in Figure 1 below.

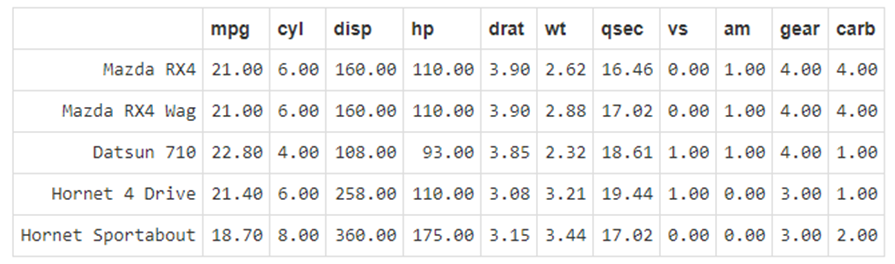


Figure 1 – mtcars dataset

Show how you would perform the following exploration and manipulation on the given dataset.

i) Lists names of variables in the dataset.

**ANSWER**

> names(mtcars)

ii) Shows the number of rows/columns of the dataset.

**ANSWER**

> dim(mtcars)

iii) See a summary statistics of the Gross horsepower.

**ANSWER**

> summary( mtcars$hp)

iv) Display the structure of the dataset.

**ANSWER**

> str(mtcars)

v) Visually check possible correlated variables.

**ANSWER**

> pairs(mtcars)

vi) Returns number of missing values in each variable of a dataset.

**ANSWER**

> colSums(is.na(mtcars))

vii) Replace all the am variables with words representation. Transmission (0 = automatic, 1 = manual)

**ANSWER**

> mtcars$am = as.character(mtcars$am)

> mtcars$am[mtcars$am==0] <- 'automatic'

> mtcars$am[mtcars$am==1] <- manual'

[8 marks]

1. Decide which descriptive statistics tool should you use according to the given situation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data type** | **Objective** | **Example** | **Statistical tools** | **Graphical tools** |
| i.  Quantitative  One variable | Estimate a frequency distribution | How many students per batch fail to graduate on time? | Frequency table | Histogram |
| ii.  Quantitative  One variable | Measure the central tendency of one sample | What is the average grade in a course? | Mean, median, mode | Box plot  Scattergram  Strip plot |
| iii.  Quantitative  One variable | Measure the dispersion of one sample | How are the grades dispersed around the average grade in a course? | Range, Standard deviation, Variance, | Box plot  Scattergram  Strip plot |
| iv.  Quantitative  Two variables | Describe the association between two variables | Does plant biomass increase or decrease with soil Pb content? | Correlation coefficients | Scatter plot  Correlation map |
| v.  Qualitative (univariate analysis) | Detect the most frequent category | Which is the most frequent eye color in Malaysia? | Mode | Bar chart  Pie chart |

[8 marks]

1. a) Explain the problems reproducibility research can and cannot solve.

**ANSWER**

* What reproducibility can solve
* Transparency
* Data Availability
* Software / Methods Availability
* Improved Transfer of Knowledge
* What reproducibility cannot solve
* Validity / Correctness of the analysis

[6 marks]

b) Distinguish between weaving and tangling in literate programming.

* **ANSWER**
  + - 1. Literate programs can be **weaved** to produce human-readable documents and **tangled** to produce machine-readable documents.

[4 marks]

3. a) Machine learning is not a solution for every type of problem. There are certain

cases where robust solutions can be developed without using ML techniques.

Discuss **TWO** situations where machine learning is useful.

**ANSWER**

ML is not needed if you can determine a target value by using simple rules, computations, or predetermined steps that can be programmed without needing any data-driven learning.

Use machine learning for the following situations:

* You cannot code the rules: Many human tasks (such as recognizing whether an email is spam or not spam) cannot be adequately solved using a simple (deterministic), rule-based solution. A large number of factors could influence the answer. When rules depend on too many factors and many of these rules overlap or need to be tuned very finely, it soon becomes difficult for a human to accurately code the rules. You can use ML to effectively solve this problem.
* You cannot scale: You might be able to manually recognize a few hundred emails and decide whether they are spam or not. However, this task becomes tedious for millions of emails. ML solutions are effective at handling large-scale problems.

[4 marks]

b) Formal ML is defined as : A computer program is said to learn from experience E, with respect to some task T, and some performance measure P, if its performance on T as measured by P improves with experience E.

Suppose your email program watches, which emails you, do or do not mark as spam, and based on that learns how to better filter spam.

What is the task **T**, experience **E** and performance measure **P** in this setting?

**ANSWER**

Spam Filter

T: Classifying emails as spam or not spam

E: Watching you label emails as spam or not spam.

P: The number (or fraction) of emails correctly classified as spam/not spam.

[3 marks]

c) How would you write a program to distinguish a picture of yourself from a picture of someone else?

**ANSWER**

Provide examples pictures of yourself and pictures of other people and let a classiﬁer learn to distinguish the two.

[3 marks]

4. The phrase “data storytelling” has been associated with many things - data visualizations, infographics, dashboards, data presentations, and so on. Too often data storytelling is interpreted as just visualizing data effectively; however, it is much more than just creating visually-appealing data charts.

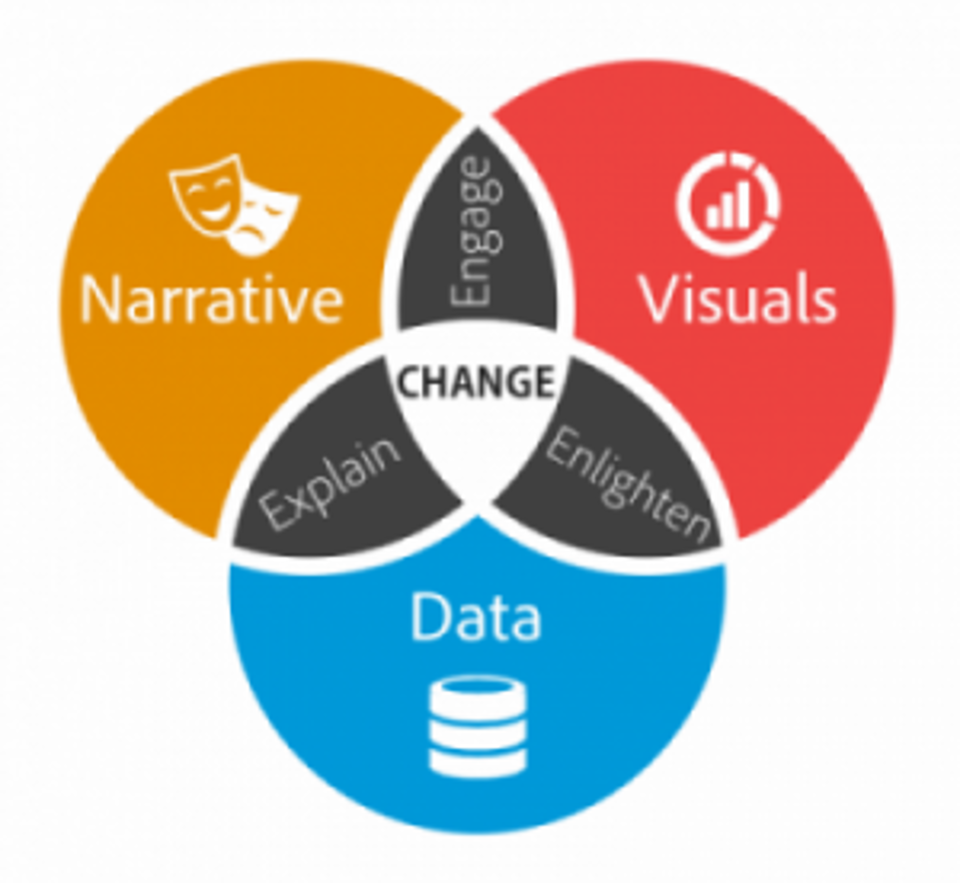
Present a **THREE** key elements structured approach of data storytelling for communicating data insights. Summarize your presentation with a diagram.

**ANSWER**

Data storytelling is a structured approach for communicating data insights, and it involves a combination of three key elements: *data*, *visuals*, and *narrative*.

It’s important to understand how these different elements combine and work together in data storytelling. When narrative is coupled with data, it helps to ***explain*** to your audience what’s happening in the data and why a particular insight is important. Ample context and commentary is often needed to fully appreciate an insight. When visuals are applied to data, they can ***enlighten*** the audience to insights that they wouldn’t see without charts or graphs. Many interesting patterns and outliers in the data would remain hidden in the rows and columns of data tables without the help of data visualizations.

Finally, when narrative and visuals are merged together, they can ***engage*** or even entertain an audience. It’s no surprise we collectively spend billions of dollars each year at the movies to immerse ourselves in different lives, worlds, and adventures. When you combine the right visuals and narrative with the right data, you have a data story that can influence and drive ***change***.



[10 marks]

**END**