MET CS 555 - Data Analysis and Visualization
Example Question

1 Hospitals Dataset

Data from n = 113 hospitals in the United States are used to assess factors related to the likelihood that a hospital patients acquires an infection while hospitalized.

Data on 12 variables:

- **Hospital:** index from 1 to 113
- **Length of stay:** average duration (in days) for all patients
- **Age:** average age (in years) for all patients
- **Infection risk:** estimated percentage of patients acquiring an infection in hospital
- **Culture:** average number of cultures for each patient without signs or symptoms of hospital-acquired infection, times 100
- **X-ray:** number of X-ray procedures divided by number of patients without signs or symptoms of pneumonia, times 100
- **Beds:** average number of beds in the hospital
- **Med school:** does the hospital have an affiliated medical school (1=Yes;2=No)
- **Region:** geographic region (1=North-East, 2=North-Central, 3=South, 4=West)
- **Patients:** average daily census of number of patients in the hospital
- **Nurses:** average number of full-time equivalent registed and licensed nurses
- **Facilities:** percent of 35 specific facilities and services which are provided by the hospital

You can read the data into R using the following commands

```
hospitals <- read.table("https://bit.ly/2qZwdMn", header =T)

# remove the first column
hospitals <- hospitals[,-1]

# Create factors variables
hospitals$Region <- as.factor(hospitals$Region)
hospitals$Med.school <- as.factor(hospitals$Med.school)
```

Questions:

1. Which single variable out of the 10 variables would you use to predict the infection risk value? Describe why you select that variables.

2. Consider your answer for previous question, consider only one single independent variable and fit a simple linear regeression (SLR) model to the data.

   Provide the equation of your SLR.

   How well your model can explain the variability of the response data?

3. Which other variables are significant predictors to be used in Multiple Linear Regression for prediciting infection risk?

4. Which other variables would you use to predicit infection risk using Multiple Linear Regression?

5. Provide a multiple linear regression equation. How well can your MLR model predict the infection risk?

6. Are the infection risks in us **Regions** different? Write your Hypothesis and provide significant tests.

7. Is number of Nurses in each hospital a significan covariate? Are the differences in different region driven by the number of Nureses?

8. Considering the case that Hospitals are affiliated medical school and they are in different reigons, are all of these hospitals different in terms of infection risks?

9. Which tuples of variables (x, y) can be used to predict the other variables? Check all possible combination and which pair has the highest model fittness? Provide reasons.

2 Credit Approval Data

This dataset is about credit card applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data.

This dataset is interesting because there is a good mix of attributes – continuous, nominal with small numbers of alues, and nominal with larger numbers of values.

You can see the data describtion here
`https://archive.ics.uci.edu/ml/datasets/Credit+Approval`
We give the variables working names based on the type of data.
'data.frame': 653 obs. of 16 variables:

- **Male**: num 1 1 0 0 0 0 1 0 0 0 ...
- **Age** : chr "58.67" "24.50" "27.83" "20.17" ...
- **Debt**: num 4.46 0.5 1.54 5.62 4 ...
- **Married** : chr "u" "u" "u" "u" ...
- **BankCustomer**: chr "g" "g" "g" "g" ...
- **EducationLevel**: chr "q" "q" "w" "w" ...
- **Ethnicity** : chr "h" "h" "v" "v" ...
- **YearsEmployed**: num 3.04 1.5 3.75 1.71 2.5 ...
- **PriorDefault** : num 1 1 1 1 1 1 1 1 1 0 ...
- **Employed** : num 1 0 1 0 0 0 0 0 0 0 ...
- **CreditScore** : num 6 0 5 0 0 0 0 0 0 0 ...
- **DriversLicense**: chr "f" "f" "t" "f" ...
- **Citizen** : chr "g" "g" "g" "s" ...
- **ZipCode** : chr "00043" "00280" "00100" "00120" ...
- **Income** : num 560 824 3 0 0 ...
- **Approved** : chr "+" "+" "+" "+" ...

```
creditAppData  <- read.csv("https://bit.ly/2HRjkxj")


head(creditAppData)
summary(creditAppData)

attach(creditAppData)
```

1. Is "Employed" a good predictor of getting credit approval? What is the c-statistic?
2. What is the risk difference between *Employed vs. Unemployed* for getting credit approval?
3. Are the varaibles Employed and Debt together good predictor for credit approval?
4. What is the risk difference for 1 unit increase in debt?
5. What is the ODD ration confidence interval for 1 unit increase in debt?
6. Is there a correlation between Income, Credit Score, and YearsEmployed and the credit approval status?