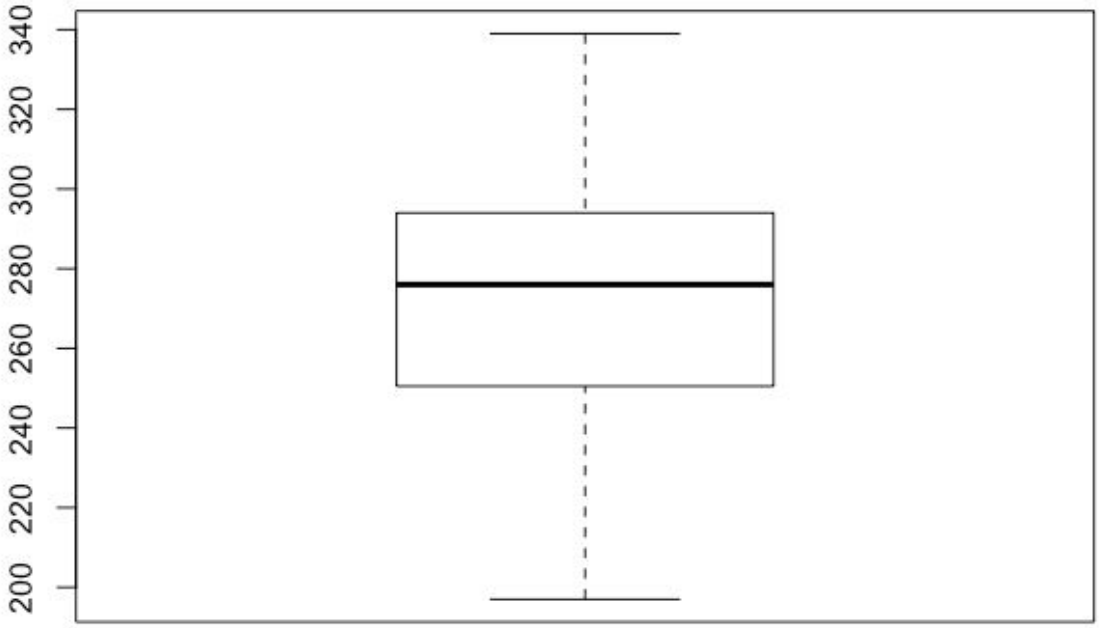
Homework 1

Ronak Mehta
SJSU ID: 014505387

Question 1

a) R code:

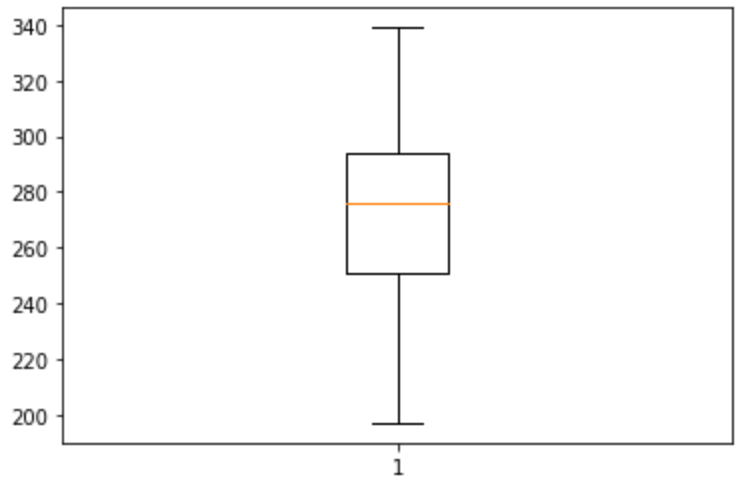
```
# A vector for the data 197, 199, 234, 267, 269, 276, 281, 289, 299, 301, and 339
A <- c(197,199,234,267,269,276,281,289,299,301,339)
# Drawing a boxplot for the above data
boxplot(A)
```



Python code:

```
#Importing plotting library 'matplotlib'
import matplotlib.pyplot as plt
#The below statement only used to plot inline in jupyter notebook
%matplotlib inline
#Given data stored in list
```

data = [197,199,234,267,269,276,281,289,299,301,339]
#The data is plotted using boxplot() function
plt.boxplot(data)



The outliers can be calculated using 1.5 * interquartile range(IQR)
In the above given data,
Median = 276
Q1 = 234
Q3 = 299
IQR = 65

Q1 - 1.5*IQR:
234 - 97.5 = 136.5

Q3+1.5*IQR:
299+1.5*65 = 396.5

Since there is no point/element below 136.5 and no point/element above 396.5. Hence there is no outlier.

---

1. b) Real-world scenarios where outliers are useful and inspiration for inquiry:
    I. Researchers in Africa discovered that some women were living with HIV totally fine for many years longer than expected despite being untreated, those rare cases were outliers compared with most untreated women, who were dying fairly rapidly. They could have been discarded as noise or error, but instead are source for inquiry.
    II. Exceptions in credit card transactions could be helpful to detect the fraudulent use of credit cards.
    III. The top 1% of people who use healthcare in the US account for over 20% of the total money spent on healthcare. These people are outliers. However, they provide us with a

tremendous amount of information about the nature of ill health and how we might better address it.

IV. In one of the images out of 386,207 images sent by curiosity rover which landed on Mars on Sol 1051, depicts a woman standing next to some type of vehicle that may be lodged on a wall.

V. The number of people who bike to work in Florida is 54,652 where as in other states its less than 16,000. In such cases it's not appropriate to exclude florida in the analysis.

---

2. a. All the below features together could possibly help in determining the veracity.

i) Features of webpage that can help determine the veracity:

1. PageRank Scores - high page rank could be considered more authentic with relatively high qualitative and quantitative links.
2. Number of false facts - A source that has few false facts could be considered trustworthy.
3. Knowledge Based Trust(KBT) estimation - Estimates the correctness of the facts and the accuracy of the sources using inference in a probabilistic model.
4. Domain - Some domain names like edu, and gov are more reliable than com, org or net since such domain could be purchased by any individual.
5. Security Trust Marks - Gives security seal to a webpage. Example - Verisign trust mark.
6. SSL certificate - Secure sockets layer digital certificate provides encrypted communication over the network which increases the trustworthiness.
7. Accolades - webpage related awards in a webpage are more trusted.
8. Fast loading web pages - Slow loading web pages are considered less secure especially while online banking.
9. Inclusion of contact details, about page, customer reviews, third party logos, social media widgets, and terms of use statements on a webpage.
10. Customer testimonials - Makes a webpage more legitimate.

ii) Features of Student Homework that can help determine the veracity:

1. Content similarity score - Score that represent the amount of similar content found in other's homework.
2. Clarity - How clear the material is.
3. Correctness - Trust increases with correct facts or answers in the homework.
4. Overwritten name or ID - A person who has overwritten his name on other's homework or not.
5. Font and format similarity - If two different homework contains similar font and format of the content.
6. Matches in the document - How much does the document's content matches with other's homework.
7. Paraphrase plagiarism - This will help even if the student has paraphrased the content.
8. Media like image screenshot similarity score - If the students use similar images for their homework, then the student's homework could be least trusted.

9. Repeated mistakes/incorrect answers similar in other's homework.

iii) Features of news item a daily that can help determine the veracity:

1. Broadcasted author - fake news often lack a source or author.
2. News material
3. Sources of the information - we can check the sources whether they were reliable in the past, do they have a bias or not. Information reported on The Wall Street Journal or New York Times are more reliable than any other little known conspiracy website.
4. Information reported on multiple sources - News limited to a single source are less reliable than information reported over various means of communication.
5. Positive results from fast-checking websites like FastCheck or The Washington's Fast Checker can spot a fake news.

iv) Features of Image on Instagram that can help determine the veracity:

1. Instagram account - verified or not
2. Number of followers - Account with more number of followers is trusted relatively more than an account with less number of followers.
3. Number of likes on that image - More number of likes could give a rough idea about the veracity of the image.
4. Number of comments on that image - More comments could represent more reliability.
5. Number of people shared the image - Since more number of people share the image, it could possibly mean that people find it trustworthy.
6. Likes per comment.
7. Account age - New account make it less reliable.
8. Data and time - This could help in validating.
9. Latitude and longitude - Cross validating with the image.
10. Device make - Could give us more information about the image.
11. Device model

v) Features of video on youtube that can help determine the veracity:

1. Verified author or not.
2. Number of subscribers
3. Number of views
4. Number of comments
5. Number of people shared the video
6. Sponsored by some company or not

---

2. b) Python code:

```
#import numpy library
import numpy as np
grades = [47,63,71,39,47,49,43,37,81,69,38,13,29,61,49,53,57,23,58,17,73,33,29]

#Find the mean and standard deviation using numpy functions mean() and std()
mean = np.mean(grades)
```

```
std = np.std(grades)
std_grade = 1/3*(std)

#Print the grades for individual score
for x in range(len(grades)):
    if grades[x]<mean-(4*std_grade):
        print("Grade for student scoring ",grades[x]," = F")
    elif grades[x]>=(mean-(4*std_grade)) and grades[x]<(mean-(3*std_grade)):
        print("Grade for student scoring ",grades[x]," = D")
    elif grades[x]>=mean-(3*std_grade) and grades[x]<mean-(2*std_grade):
        print("Grade for student scoring ",grades[x]," = C-")
    elif grades[x]>=mean-(2*std_grade) and grades[x]<mean-std_grade:
        print("Grade for student scoring ",grades[x]," = C")
    elif grades[x]>=mean-std_grade and grades[x]<mean:
        print("Grade for student scoring ",grades[x]," = C+")
    elif grades[x]>=mean and grades[x]<mean+std_grade:
        print("Grade for student scoring ",grades[x]," = B-")
    elif grades[x]>=mean+std_grade and grades[x]<mean+(2*std_grade):
        print("Grade for student scoring ",grades[x]," = B")
    elif grades[x]>=mean+(2*std_grade) and grades[x]<mean+(3*std_grade):
        print("Grade for student scoring ",grades[x]," = B+")
    elif grades[x]>=mean+(3*std_grade) and grades[x]<mean+(4*std_grade):
        print("Grade for student scoring ",grades[x]," = A-")
    elif grades[x]>=mean+(4*std_grade) and grades[x]<mean+(5*std_grade):
        print("Grade for student scoring ",grades[x]," = A")
    else:
        print("Grade for student scoring ",grades[x]," = A+")

#Output
Grade for student scoring  47  = B-
Grade for student scoring  63  = B+
Grade for student scoring  71  = A
Grade for student scoring  39  = C
Grade for student scoring  47  = B-
Grade for student scoring  49  = B-
Grade for student scoring  43  = C+
Grade for student scoring  37  = C
Grade for student scoring  81  = A+
Grade for student scoring  69  = A-
Grade for student scoring  38  = C
Grade for student scoring  13  = F
Grade for student scoring  29  = D
Grade for student scoring  61  = B+
Grade for student scoring  49  = B-
Grade for student scoring  53  = B
Grade for student scoring  57  = B
Grade for student scoring  23  = F
Grade for student scoring  58  = B
```

Grade for student scoring 17 = F
Grade for student scoring 73 = A
Grade for student scoring 33 = C-
Grade for student scoring 29 = D

---

3. a) Some examples of how Social Media and Big Data in general is helping with the problem of veracity:

i) Veracity and velocity of social media content on breaking news:

- Since 65% of users receive breaking news through social media and has become the main source of breaking news online with 2.4 billion users, it's crucial to determine the veracity of the news.
- Journalists verify the veracity of the event by contacting the social media user who took the photo of the event before broadcasting their news channel.
- They use a crawler to get information from various Social Media APIs like twitter, youtube, instagram then perform tokenization, entity extraction, and temporal segmentation to get a trust model which results in TrustedScore and TrustedContent.

ii) Posts on Facebook market or Craigslist -

- Social media reviews and in general big data can be used to find out about the products sold on Facebook market or craigslist.
- Analysis like can find out about the veracity of the posts, seller's previous sold items, seller's reputation and location.

iii) Social media changed Finance with the veracity problem -

- With the advent of Big Data social media streaming data are used to better find the target customers.
- While evaluating a new investment fund if investors are concerned about the risk then investors can take the help of social media and big data to investigate the good, bad or ugly from other investors and analysis on finance data.

---

3. b) Five recent news item which demonstrate how Machine Learning is being used to advance the society:

1. Self-driving cars and automated transportation -
    - The self-driving cars works on machine learning algorithms which take the input features (like real-time visual, sensor data) and gives an output, decision among possible next actions of a car with little or no human input.
    - Distracted or aggressive driving and traffic collisions will be reduced with the autonomous vehicle.
    - It would reduce the labor costs, relieve travellers from navigation chores and increase leisure time.
    - Improve the fuel economy of the car by optimizing the drive cycle.
    - Autonomous vehicles can reduce the need for parking space.

2. Benefits in healthcare -
    - Google developed a machine learning algorithm to detect cancerous tumors on mammograms.
    - Stanford used a deep learning algorithm to identify skin cancer.

- The Journal of the American Medical Association article reported the results of a deep machine-learning algorithm that diagnosed diabetic retinopathy in retinal images.
- Microsoft's InnerEye used machine learning to differentiate between tumors and healthy anatomy using 3D radiological images that aided medical experts in radiotherapy and surgical planning.

3. IBM’s Green Horizon Project -
    - It analyzes environmental data from thousands of sensors and sources, to give accurate and evolving weather forecasts.
    - It model ways to mitigate environmental impact and supports cleaner air and increase the use of renewable energy.

4. Drones and machine learning technology are used to perform dangerous tasks such as bomb disposal and welding.
    - This saved thousands of lives.
5. Online transportation networks -
    - In order to minimize the detours, transportation companies like Uber ATC revealed that they use machine learning algorithms to define price surge hours by predicting the rider demand.
6. Face recognition -
    - Face recognition is used by Apple and Facebook to categorize the pictures uploaded in the media and tag it a name by using Machine learning models with the previous images.

4.
```
#Import the necessary libraries
import pandas as pd
from sklearn.utils import resample
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split

#Read the file using read_csv()
bc = pd.read_csv("/Users/ronakmehta/Desktop/BreastCancer.csv",encoding="utf-8")
bc.columns

#Remove 'Unnamed: 0' and 'Id' columns
bc = bc.drop(['Unnamed: 0', 'Id'],axis=1)

#The 'bc' dataframe does not contain any null values
bc.isnull().sum()

bc.Class.unique()
# Convert the Class categorical variable into numeric
bc['Class'] = bc['Class'].map({'malignant':1, 'benign':0})
bc.head()
bc.dtypes
bc['Class'].value_counts()
```

```
X_train, X_test, y_train, y_test = train_test_split(bc,
        bc['Class'], test_size=0.30,
         random_state=100)

# Balance the Data
bc_majority = X_train[X_train.Class==0]
bc_minority = X_train[X_train.Class==1]
X_train.Class.value_counts()
bc_majority_downsampled = resample(bc_majority, replace=False, n_samples= 164,
random_state=100)
bc_downsampled = pd.concat([bc_majority_downsampled,bc_minority])
bc_downsampled.Class.value_counts()
y = bc_downsampled.Class
x = bc_downsampled.drop('Class',axis=1)

# Implement Logistic Regression model
#Please don't add the solver parameter on your kernel or IDLE
logitMod = LogisticRegression(solver='lbfgs')
logitMod.fit(x,y)

# Predict the values with the trained model
X_test = X_test.drop('Class',axis=1)
Predictions = logitMod.predict(X_test)

# Obtain the accuracy
score = logitMod.score(X_test,y_test)
print(score)
```

---

5. a) Other mathematical functions that can possibly take the place of the Sigmoid function and help with Machine Learning:
   i) tanh or hyperbolic tangent activation function -

$$f(x) = \tanh(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$$

   - TanH has a range from -1 to 1 and is similar to logistic sigmoid having s-shaped curve.
   - Easier to model inputs with strongly negative, neutral and positive values.
   - The function is monotonic and works better than sigmoid.

   ii) Rectified Linear Unit(ReLU) -

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ x & \text{for } x > 0 \end{cases}$$

   - Range is from zero to $+\infty$.
   - It allows the network to converge and also allows for backpropagation.

- But any negative input values becomes zero which decreases the ability to fit the model with the data properly.

iii) Leaky ReLU -

$$f(x) = \begin{cases} 0.01x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$$

- The range is from (0.01, $+\infty$).
- Enables back-propagation even for negative input values.
- It's used in the case of dead neurons.

iv) S shaped Rectified Linear Activation unit(SReLU) -

$$f_{t_l,a_l,t_r,a_r}(x) = \begin{cases} t_l + a_l(x - t_l) & \text{for } x \leq t_l \\ x & \text{for } t_l < x < t_r \\ t_r + a_r(x - t_r) & \text{for } x \geq t_r \end{cases}$$

t_l, a_l, t_r, a_r are parameters.

- The range is from ($-\infty$, $+\infty$).
- The function is not monotonic.
- Gives sparsity, the more units that exist in a layer the more sparse the resulting representation. Sigmoids on the other hand are always likely to generate some non-zero value resulting in dense representations.
- Sparse representations seem to be more beneficial than dense representations

v) Sinc -

$$f(x) = \begin{cases} 1 & \text{for } x = 0 \\ \frac{\sin(x)}{x} & \text{for } x \neq 0 \end{cases}$$

- The range is from [-.217234,1]
- The inputs gets easily converted to zero.

vi) Gaussian

$$f(x) = e^{-x^2}$$

- The range is from (0,1].

---

5. b. i) A coin classification system for a vending machine given weight, size and denomination to classify coins doesn't require any machine learning to classify coins.

ii) If Video surveillance camera feeds contain previous labeled data, then violence can be detected using semi-supervised learning since the violence could be handled with various new weapons which did not exist in previous records or the violence act could be new.
iii) Since the goal is to find out the disease re-emergence with past data and current data, supervised learning can be used to help us get the outcome. Because past and current data contain the labels to appropriately use supervised learning and predict the re-emergence of the disease.
iv) Unsupervised learning could be used to identify new plant diseases based on leaf images. Since there are no records with such a disease, unsupervised learning is the best option. If the motive is to find out the category of the disease it belongs to then semi-supervised learning could be used to identify the new disease.
v) In strategized chess-playing since the motive is to increase the rewards by adjusting the strategy so as to penalize that leads to losing, reinforcement learning could be used to help predict the actions based on penalty.