ECE 657A

Homework 1

Data Summarization

Abul Hasan Fahad (#20764979)

10-Jan-19

Solution:

Dataset: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Code:

```
#https://pandas.pydata.org/pandas-docs/stable/api.html#dataframe
#https://seaborn.pydata.org/index.html
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style="darkgrid")

from pandas import set_option
df=pd.read_csv('E:\ECE 657A\winter 2019\homework\data.csv')

shape = df.shape
print(shape)

types = df.dtypes
print(types)

set_option('display.width', 100)
set_option('precision', 3)

description = df.describe()
print(description)

mode = df.mode()
print(mode)

variance= df.var()
print(variance)

skew=df.skew()
print(skew)

kurt=df.kurtosis()
print(kurt)

PCC= df.corr(method='pearson', min_periods=1)
print (PCC)


writer = pd.ExcelWriter('output.xlsx')
description.to_excel(writer,'Sheet1')
variance.to_excel(writer,'Sheet2')
skew.to_excel(writer,'Sheet3')
kurt.to_excel(writer,'Sheet4')
PCC.to_excel(writer,'Sheet5')
mode.to_excel(writer,'Sheet6')
writer.save()

sns.heatmap(PCC,center=0,cmap="YlGnBu")
```

Question 1: In the cancer dataset report the mean, mode and skew, standard deviation and variance values for all the continuous valued features.

Answer:

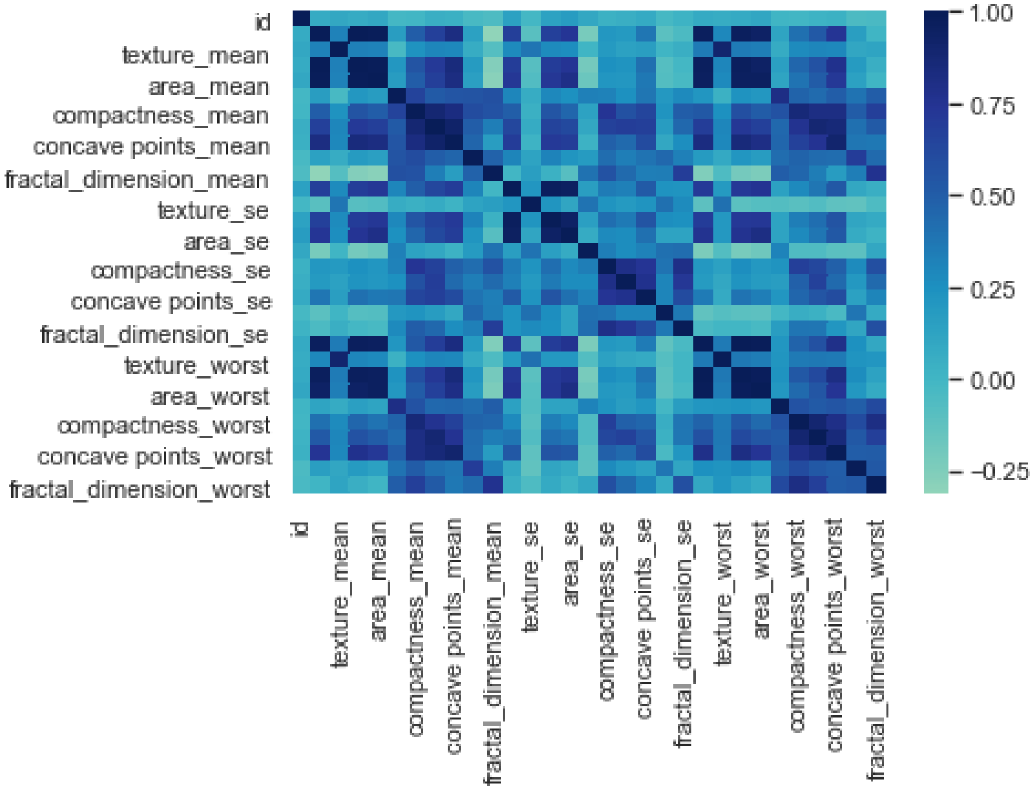
a) Mean: Sheet 1 of “output.xlsx”
b) Mode: Sheet 6 of “output.xlsx”
c) Skew: Sheet 3 of “output.xlsx”
d) Standard deviation: Sheet 1 of “output.xlsx”
e) Variance: Sheet 2 of “output.xlsx”

Additional: Kurtosis: Sheet4 in “output.xlsx”

Question 2: In the cancer dataset s a few pairs of features for correlation by computing their PCC and report the resulting numbers and explain what they mean.

Answer:

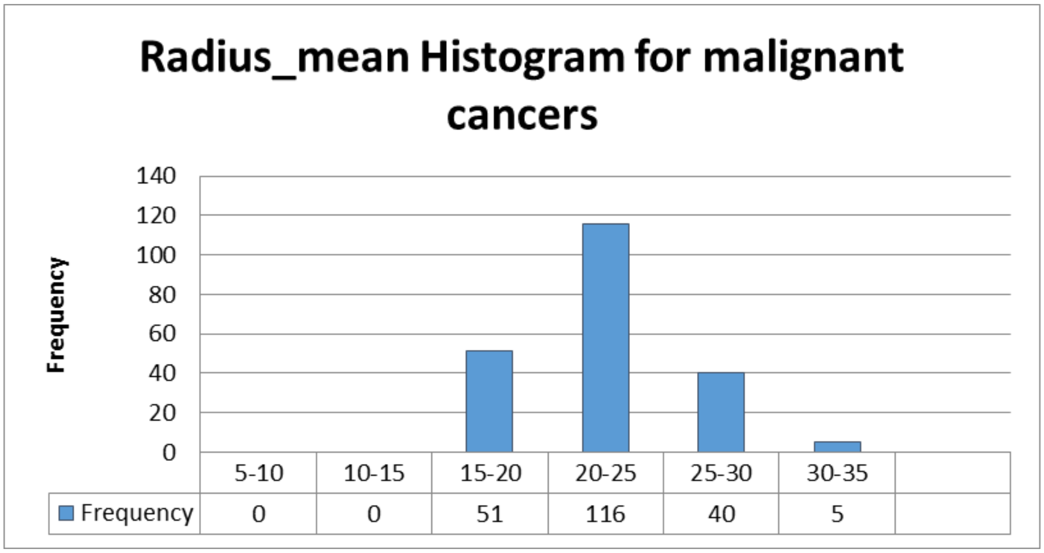
Pearson Correlation Matrix can be found in Sheet 5 of “Output.xlsx”. Heatmap of the matrix can be found below. The correlation coefficient value is depicted by the colorbar.



To explain, an example can be given. From the result, it can be concluded that texture_mean shows positive correlation with radius_mean, perimeter_mean, area_mean, compactness_mean, concavity_mean, concave points_mean and symmetry_mean. It shows negative correlation with smoothness_mean and fractal_dimension_mean.

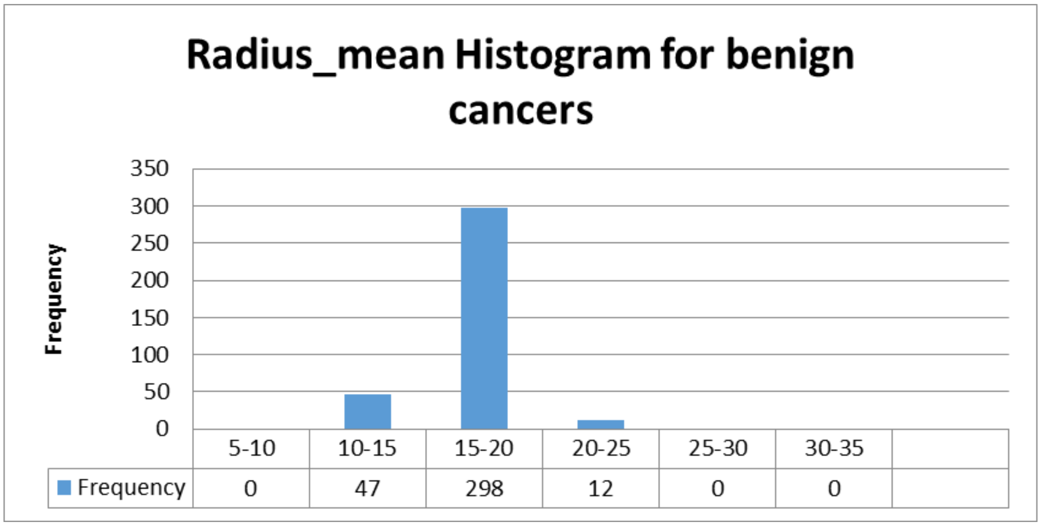
Question 3: In the cancer dataset plot two histograms for a continuous valued feature of your choice: One for patients with each diagnosis (M or B).

Answer: Radius_mean attribute was selected for the analysis. Data was filtered as per diagnosis (M/B) and gathered in separate columns. Excel Data Analysis toolpack was used to generate Histograms which are given below:



Radius_mean Histogram for malignant cancers

	5-10	10-15	15-20	20-25	25-30	30-35	
■ Frequency	0	0	51	116	40	5	



Radius_mean Histogram for benign cancers

	5-10	10-15	15-20	20-25	25-30	30-35	
■ Frequency	0	47	298	12	0	0	