Data analysis and visualisations

Name of student

Name of professor

University

Course

Date

The analysis was done using Python and visualizations done on Jupyer notebook. The data used for the assignment was manually collected and recorded. The data consists of the number of users who went to the nearby shop within the nearest kilometer and had their counts taken. The column variables consists of the count of shop walk-ins against the sum of each and every hourly walk-in to the store.

The first step is to load the data into the Notebook and read it as below:

#Load the dataset and read it on Jupyter

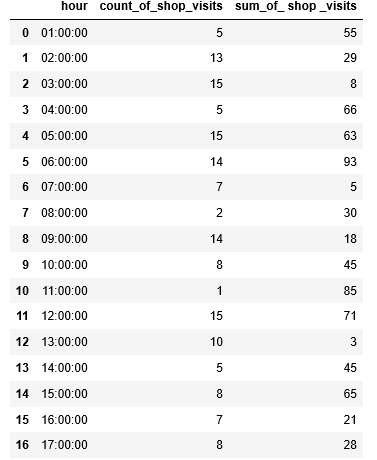
from pandas.io.json import json\_normalize

import json

import pandas as pd

dataset = pd.read\_csv("shop\_visits.csv")

dataset



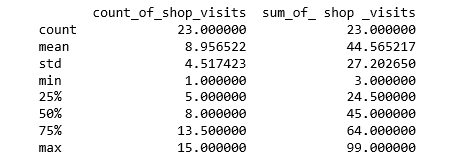
Getting a summary of the dataset looks like below:

#Get the summary of the data

df = pd.DataFrame(dataset)

print(df)

print(df.describe())



The summary of the dataset indicates the mean, std deviation, minimum 50th percentile, 75th percentile and maximum value. The next step is to visualize the dataset on Notebook.

# creating some arrays relevant for the plot

the\_hour = ['1', '2', '3','4','5','6', '7', '8','9','10',

            '11', '12', '13','14','15','16', '17', '18','19','20','21', '22', '23']

x\_pos = np.arange(len(the\_hour))

CTEs = [shop\_walkin,total\_visits]

#Building the plot

fig, ax = plt.subplots()

ax.bar(x\_pos, CTEs, align='center', alpha=0.5)

ax.set\_ylabel('Hourly')

ax.set\_xticks(x\_pos)

ax.set\_xticklabels(the\_hour)

ax.set\_title('Shop walkin')

ax.yaxis.grid(True)

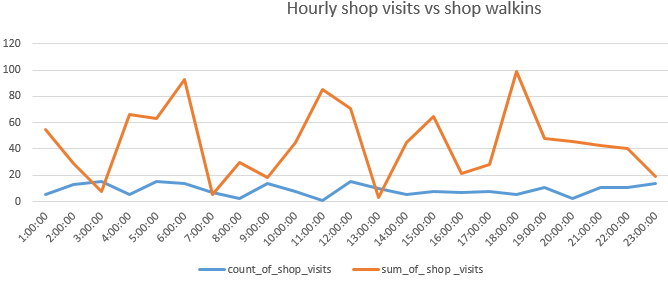
# Saving result as an image

plt.tight\_layout()

plt.savefig('bar-chart.png')

plt.show()

The graph generated looks like below. This indicates that most shop visits happened at 5am-6am, 11am and 6pm. Most likely contribution to this could be alluded to the fact that these are the times when people were going to buy food.



Further, a probability prediction analytics was done on the datasets using the regression model from a possible sample population of 400 and then a predictive case done to determine how many people would most likely walk in to the shop at 2pm and 4pm and the output returned 259 possible persons. This is 60% true based on the fact that there is a positive correlation between the times when people go to buy food and the sum of persons going to the store. So the linear regression model is correct.

#Probabilistic predictions here:

from sklearn.linear\_model import LinearRegression

from sklearn.datasets import make\_regression

# Next step is to generate regression dataset

X, y = make\_regression(n\_samples=400, n\_features=2, noise=0.1)

# then fit the model as required

model = LinearRegression()

model.fit(X, y)

# create a new data instance

Xnew = [[2,4]]

# the predict the outcome

ynew = model.predict(Xnew)

# show prediction

print("X=%s, Predicted=%s" % (Xnew[0], ynew[0]))

