**CA 614 – Data Mining and Analytics**

**Assignment – II**

**Data Pre-processing Tasks – Part II**

1. **Binning using Python**

import pandas as pd

df = pd.read\_csv("D:\\20BSIT097\\Automobile\_data.csv")

print(df)

* **Implement Binning using cut and qcut methods.**

#cut

df['wheel-base'].min()

df['wheel-base'].max()

bins=[88,99,120]

labels=['small','medium']

df['grade']=pd.cut(df['wheel-base'],bins=bins,labels=labels,include\_lowest=True)

print(df['grade'],df['wheel-base'])

df['grade'].value\_counts()

#qcut

df['grades']=pd.qcut(df['wheel-base'],q=2)

print(df['grades'])

df['grades'].value\_counts()

* **Also, transform the bins values.**

#transform the bins value

df.groupby('grades')['wheel-base'].transform('mean')

1. **Outlier detection and removal**

* **Detect the outlier using visualization method**

#outlier

import pandas as pd

import numpy as np

na1=["n.a","not available"]

df=pd.read\_csv("D:\\20BSIT097\\Automobile\_data.csv",na\_values=na1)

print(df)

#boxplot

import seaborn as sns

sns.boxplot(x=df['price'])

#outlier

import pandas as pd

import numpy as np

na1=["n.a","not available"]

df=pd.read\_csv("D:\\20BSIT097\\Automobile\_data.csv",na\_values=na1)

print(df)

#scatter plot

import matplotlib.pyplot as plt

fig,ax=plt.subplots(figsize=(18,10))

ax.scatter(x=df['price'],y=df['num-of-cylinders'])

plt.show()

#histogram

df.hist(column='average-mileage',bins=10)

df.hist()

* **Detect the outlier using statistical method**

#statistics of outlier

from scipy import stats

import numpy as np

df['price z scores'] = stats.zscore(df['price'],nan\_policy='omit')

print(df.head(40))

#position of the outlier

thresold = 3

print(np.where(df['price z scores']>2))

* **Treat the outliers**

#treat the outlier

Q1 = df.quantile(0.25)

Q3 = df.quantile(0.75)

IQR = Q3 - Q1

print(IQR)

Q1 = df['horsepower'].quantile(0.25)

Q3 = df['horsepower'].quantile(0.75)

IQR = Q3 - Q1

print(IQR)

upper = Q3 + 1.5 \* IQR

lower = Q1 - 1.5 \* IQR

print(upper)

print(lower)