**CA 614 – Data Mining and Analytics**

**Assignment – I**

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

1. **Get an overview of the dataset**

* **Read the .csv file**

import pandas as pd

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

* **Read .csv file and consider “na”, “N/A”, “Not Available” as set of missing values**

na1= ["na.","N/A","Not Available"]

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

print(df)

* **Print the entire dataset in Pandas data frame**

print(df)

* **Print datatype of all the columns of dataset**

df.info()

1. **Get a statistical summary of the dataset**

* **Display all the statistical indicators of your dataset**

df.describe()

* **Print the name of all unique columns**

print(df["wheel-base"].nunique)

print(df["company"].unique)

print(df["engine-type"].unique)

* **Print the count of all unique columns**

print(df.nunique())

1. **Get a subset of the entire dataset**

* **Create a new Pandas dataframe and copy ith rows and jth columns of your dataset**

df1 = df.iloc[5:15]

print(df1)

1. **Modify the dataset**

df.rename(columns = {'wheel-base':'wheel'})

df.columns.str.upper()

1. **Identify and deal with missing values**

* **Display all the tuples with status of missing values**

df.isnull()

* **Count the number of missing values in dataset of columns**

df.isnull().sum()

* **Print the columns having missing value**

n1 = pd.isnull(df['price'])

df[n1]

* **Drop all the rows having atleast one missing values in the same Pandas dataframe**

new\_df = df.dropna(axis = 0, how ='any')

new\_df.info()

* **Drop all the rows where all the values are missing values in the same Pandas dataframe**

new\_df= df.dropna(axis=0, how='all')

new\_df.info()

* **Drop all the columns having atleast one missing values using different Pandas dataframe**

newdf=df.dropna(axis=1,how='any')

newdf

* **Drop all the columns where all the values are missing values using different Pandas dataframe**

newdf=df.dropna(axis=1,how='all')

newdf

* **Fill the missing value with constant**

df1=df.fillna('0')

df1.info()

* **Fill the missing value with mean, median and mode**

meandf=df['price'].fillna(df['price'].mean())

meandf.iloc[21:25]

newdf=df['price'].fillna(df['price'].median())

newdf.iloc[21:25]

newdf1=df['price'].fillna(df['price'].mode())

newdf1.iloc[21:25]

* **Fill the missing value with forward fill, backward fill and interpolate**

dfnew=df.fillna(method='ffill')

dfnew.iloc[19:25]

dfnew1=df.fillna(method='bfill')

dfnew1.iloc[19:26]

df.interpolate(method ='linear', inplace=True)

df.iloc[21:25]