|  |  |
| --- | --- |
|  | 1. Various data exploration operations and data cleaning using python |

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Student** | **Hardik Prajapati** | **Roll No.** | **9152** |
| **Sign here to indicate that you have read all relevant material provided /available on Moodle while performing and writing this experiment** | | **Sign:** | |

**Late Submission Details (if any)**

|  |  |  |
| --- | --- | --- |
| **Reason(s) of late submission** | **Date of practical performance** | **Date of practical submission** |
|  |  |  |

**References used**

|  |  |  |
| --- | --- | --- |
| 1 | Name and author of reference book(s) with page nos. |  |
| 2 | Name and roll nos. of the peers whose help you have taken (if any) |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Rubrics for assessment of Experiment:**   |  |  |  |  | | --- | --- | --- | --- | | Indicator | Poor | Average | Good | | Timeliness  Maintains Experiment deadline (3) | Experiment not done (0) | One or More than One week late (1-2) | Maintains deadline (3) | | Completeness and neatness  Complete all parts of Experiment (3) | N/A | < 80% complete (1-2) | 100% complete (3) | | Originality  Extent of plagiarism (2) | Copied it from someone else (0) | At least try to implement but could not succeed (1) | Implemented (2) | | Knowledge  In depth knowledge of the Experiment (2) | Unable to answer any questions (0) | Unable to answer few questions (1) | Able to answer all questions (2) | |
| **Assessment Marks:**   |  |  | | --- | --- | | Timeliness |  | | Completeness and neatness |  | | Originality |  | | Knowledge |  | | Total |  | |

**Signature of Teacher with date**

|  |  |
| --- | --- |
| **1.** | **Course, Subject & Experiment Details** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Course & Branch** | **T.E. (ECS)** | **Estimated Time** | **02 Hours Per Week** |
| **Current Semester** | **Semester V** | **Subject Name** | **DWM** |
| **Chapter No. & Unit** | **2** | **Chapter Title** | **ETL Operations** |
| **Experiment Type** | **Software Performance** | **Subject Code** | **ECC 604** |

|  |  |
| --- | --- |
| **2.** | **Aim & Objective of Experiment** |

1. Understand various data exploration and data cleaning operations ( CO 603.3 )

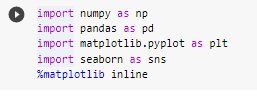
|  |  |
| --- | --- |
| **3.** | **Expected Outcome of Experiment** |

1. To understand various data exploration operations and techniques such central tendency, standard deviation, mean, median, correlation etc.

|  |  |
| --- | --- |
| **4.** | **Code of the experiment** |

**Implementation:**

1. Load the libraries

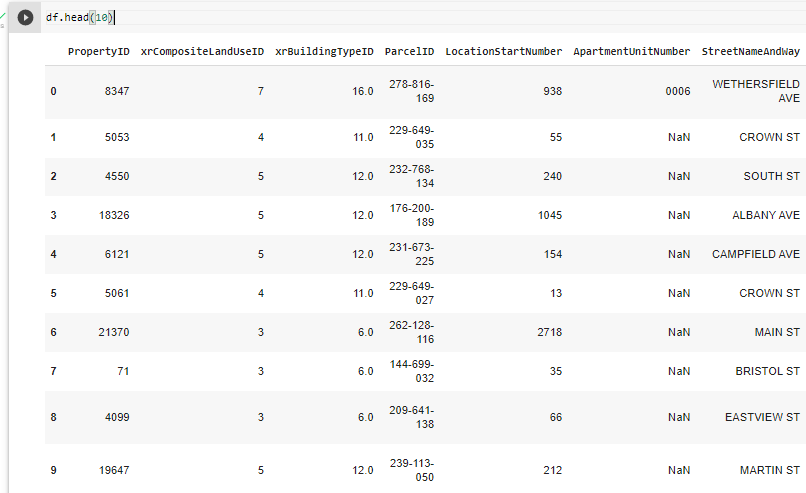


2) Download the data set from <https://data.hartford.gov/Financial/Real-Estate-Sales-730-Days/3cs2-w2b6/data>

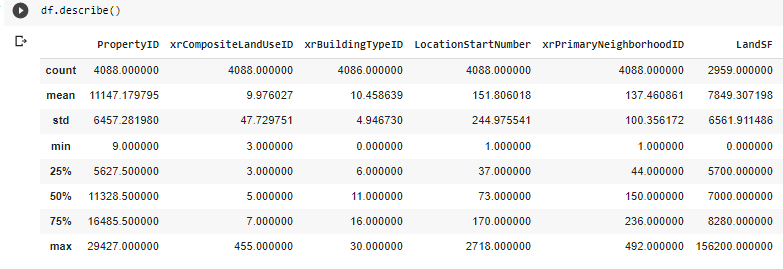
3) Read the file –select appropriate file read function according to data type of file.



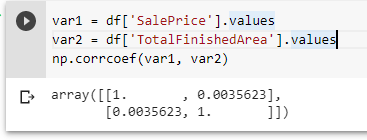
4) Display attributes in the data set-10 samples.



5) Describe the attributes name, count no of values, and find min, max, mean data type, range, quartile, and percentile.



6) Display correlation between any two attributes.



7) Download the csv file which contains missing data

8) Identify missing values fill them using appropriate technique using Python



1. **Paste your python code along with the output and screen shots**

*from* google.colab *import* files

uploaded = files.upload()

*# Commented out IPython magic to ensure Python compatibility.*

*import* numpy *as* np

*import* pandas *as* pd

*import* matplotlib.pyplot *as* plt

*# %matplotlib inline*

df = pd.read\_csv("Real\_Estate\_Sales\_730\_Days.csv")

df.head(10)

df.describe()

df = df.drop(['ApartmentUnitNumber'],*axis*=1)

land\_sf\_mean = df['LandSF'].mean()

df['LandSF'].fillna(*value* = land\_sf\_mean , *inplace* = True)

Total\_finish\_sf\_mean = df['TotalFinishedArea'].mean()

df['TotalFinishedArea'].fillna(*value* = Total\_finish\_sf\_mean , *inplace* = True)

fillname = 'ALDWIN '

df['OwnerFirstName'].fillna(*value* = fillname , *inplace* = True)

fillnme = 'MCKELLAR'

df['OwnerLastName'].fillna(*value* = fillname , *inplace* = True)

df = df.dropna()

df.info()

df.isnull().sum()

df.describe()

df.duplicated().sum()

df.to\_csv('Filtered\_Real\_Estate\_Sales\_730\_Days.csv',*encoding*='utf-8-sig')

files.download('Filtered\_Real\_Estate\_Sales\_730\_Days.csv')

uploaded = files.upload()

df = pd.read\_csv('Filtered\_Real\_Estate\_Sales\_730\_Days.csv')

df.head()

y = df['SalePrice'].values

x = df['LandSF'].values

df.plot(*x*='SalePrice',*y*='LandSF',*style*='p')

plt.xlabel('Sale Price')

plt.ylabel('Land Square Foot')

plt.show()

sales\_sub\_cats = df.groupby('AssrLandUse')['TotalFinishedArea'].agg(['sum'])

plt.style.use('bmh')

sales\_sub\_cats.plot(*xlabel* ='Land Use', *ylabel* = "Finished Area",*kind* = 'bar',*figsize*=(10,10))

plt.show()

region = df['AssrLandUse'].value\_counts().index

x = df['AssrLandUse'].value\_counts()

plt.figure(*figsize*=(15,15))

plt.pie(x,*explode* = None , *labels* =  region ,*autopct*='%2.2f%%')

plt.title("Sales of each category")

plt.show()

y = df['SalePrice'].values

numeric = df.\_get\_numeric\_data()

x = numeric.drop(['SalePrice','Unnamed: 0','PropertyID','xrCompositeLandUseID','xrBuildingTypeID','LocationStartNumber','xrPrimaryNeighborhoodID','xrSalesValidityID','xrDeedID'],*axis*=1).values

*from* sklearn.model\_selection *import* train\_test\_split

x\_train, x\_test, y\_train, y\_test = train\_test\_split( x, y, *test\_size*=0.2, *random\_state*=69)

*from* sklearn *import* linear\_model

reg = linear\_model.LinearRegression()

reg.fit(x\_train,y\_train)

y\_pred = reg.predict(x\_test[1:3])

print(y\_pred)

*from* sklearn.metrics *import* r2\_score

e = r2\_score(y\_test,y\_pred)

e

*from* sklearn.metrics *import* mean\_squared\_error

e = mean\_squared\_error(y\_test,y\_pred)

e

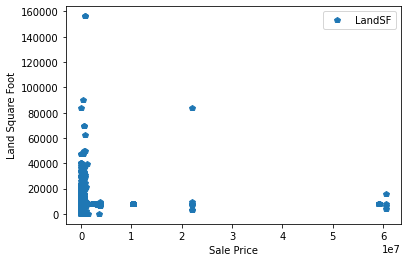
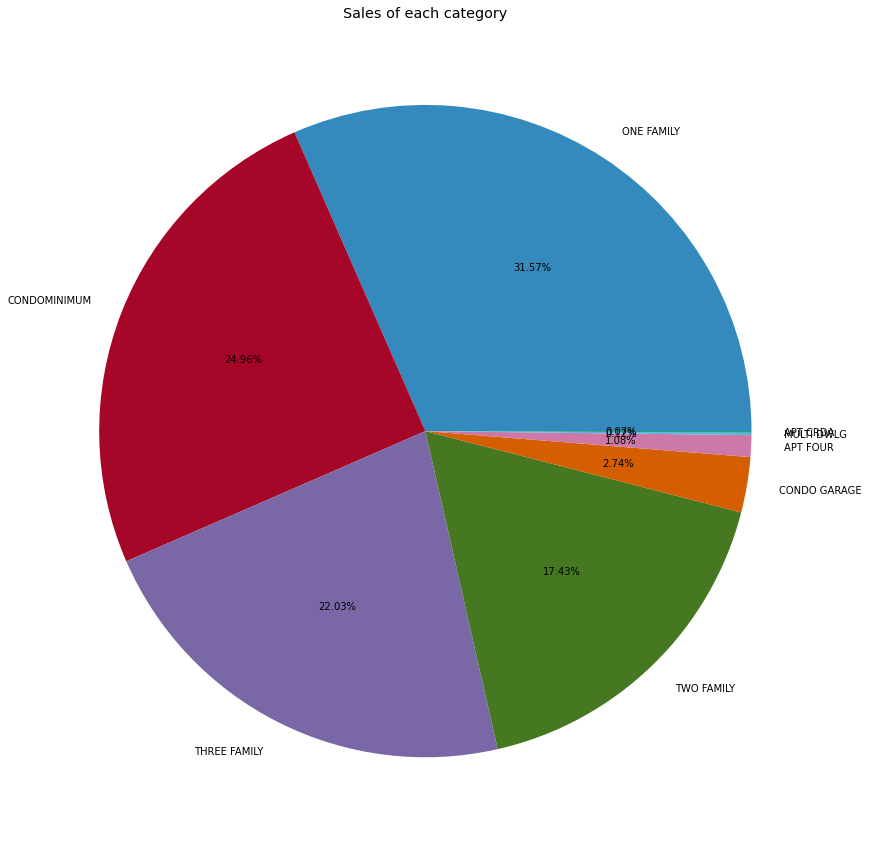
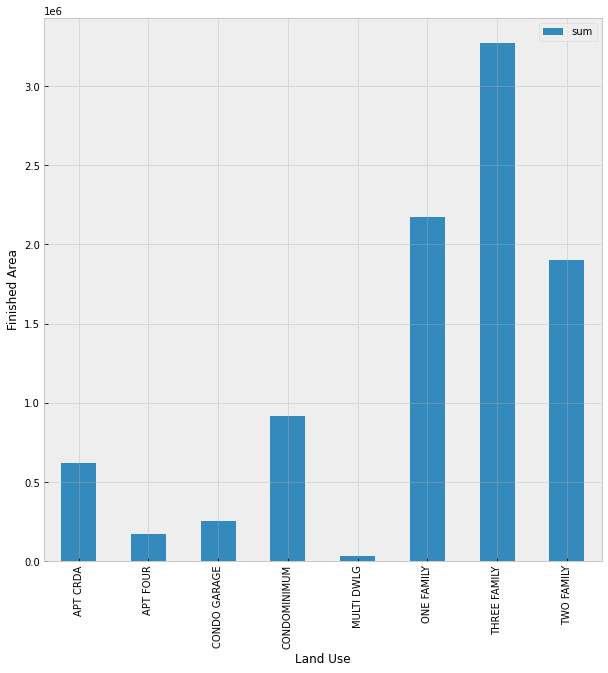
|  |  |
| --- | --- |
| **5.** | **Conclusions & Inferences** |

Through this experiment I was able to visualize the data through graphs and plots. And also I filtered the

given data.

|  |  |
| --- | --- |
| **6.** | **Post Lab exercise** |

1. Give visualization of statistical description of data – in form of histogram, scatter plot and box plot.



1. Find the linear regression for the given dataset and paste the python code along with the output.

