Cleaning Data in Python

November-15-17 10:38 PM

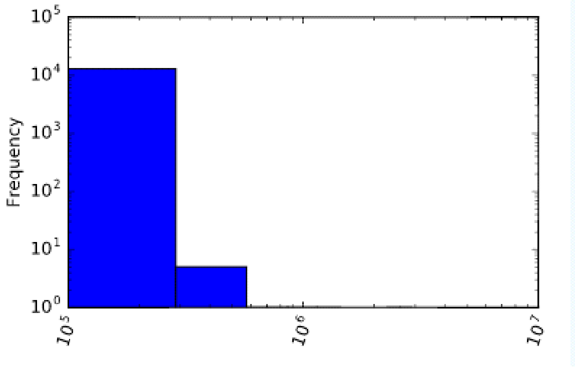
https://campus.datacamp.com/courses/cleaning-data-in-python/exploring-your-data?ex=6

```
#Ch 1 Exploring your data
# Import pandas
import pandas as pd

# Read the file into a DataFrame: df
df = pd.read_csv ('dob_job_application_filings_subset.csv')

# Print the head of df
print(df.head())

# Print the tail of df
print(df.tail())

# Print the shape of df
print(df.shape)

# Print the columns of df
print(df.columns)

# Print the head and tail of df_subset
print(df_subset.head())
print(df_subset.tail())
```

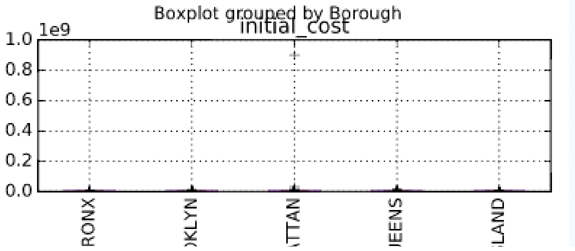
```
#Further diagnosis
# Print the info of df
print(df.info())

# Print the info of df_subset
print(df_subset.info())
```

```
#Frequency counts for categorical data
# Print the value counts for 'Borough'
print(df['Borough'].value_counts(dropna=False))

# Print the value_counts for 'State'
print(df['State'].value_counts(dropna=False))

# Print the value counts for 'Site Fill'
print(df['Site Fill'].value_counts(dropna=False))
```

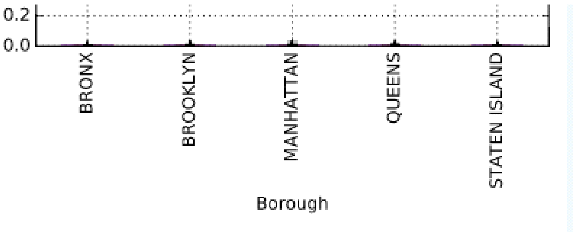
```
#Visualizing single variables with histograms
# Import matplotlib.pyplot
import matplotlib.pyplot as plt

# Plot the histogram
df['Existing Zoning Sqft'].plot(kind='hist', rot=70, logx=True, logy=True)

# Display the histogram
plt.show()
```



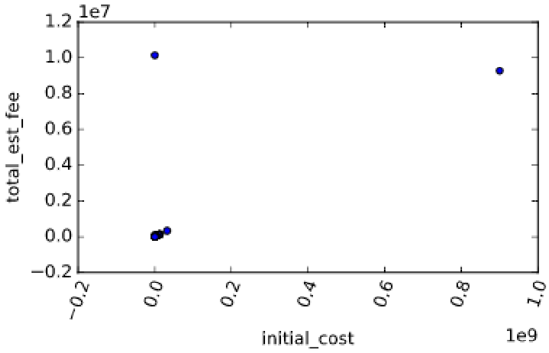
```
#Visualizing multiple variables with boxplots
# Import necessary modules
import pandas as pd
import matplotlib.pyplot as plt

# Create the boxplot
df.boxplot(column='initial_cost', by='Borough', rot=90)

# Display the plot
```

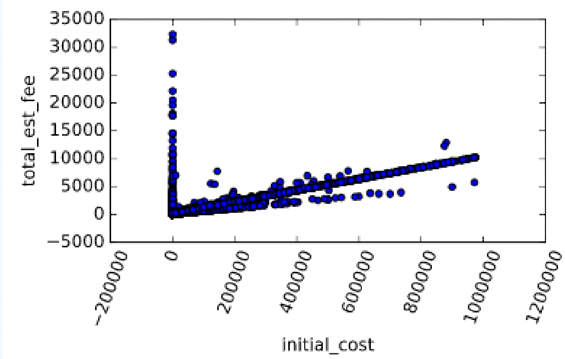


```
# Create the boxplot
df.boxplot(column='initial_cost', by='Borough', rot=90)

# Display the plot
plt.show()
```



```
#Visualizing multiple variables with scatter plots
# Import necessary modules
import pandas as pd
import matplotlib.pyplot as plt

# Create and display the first scatter plot
df.plot(kind='scatter', x='initial_cost', y='total_est_fee', rot=70)
plt.show()

# Create and display the second scatter plot
df_subset.plot(kind='scatter', x='initial_cost', y='total_est_fee', rot=70)
plt.show()
```


1e7
total_est_fee
initial_cost
1e9


total_est_fee
initial_cost