Project Report

Project Title: Ontario Rental Market Data Analysis **Using Machine Learning**

Project Phase: Exploratory Data Analysis (EDA) and Visualization).

Reporting Period: January 2024-April 2024

Project Overview: The project involved conducting Exploratory Data Analysis (EDA) on a dataset obtained from Kijiji, focusing on small community rental properties in Ontario. The dataset comprised various features such as the number of bedrooms, types of properties, size, and amenities that affect rental marketing.

PART-I

Accomplishments:

EDA and Data Visualization

Data Collection and Cleaning:

- Obtained the dataset from Kijiji, containing information about rental properties in Ontario having 25732 rows and 18 columns.
- Conducted initial data exploration to understand the structure and content of the dataset. Below are details about the columns that we have in our data:
    1. adId: A unique identifier assigned to each rental property listing, used for tracking, and referencing individual listings.
    2. CSDUID: A unique identification number associated with each rental property, potentially used for administrative or database purposes.
    3. CSDNAME: Name of the property.
    4. Title: The descriptive title or headline associated with the rental property listing, providing a summary of its key features or attractions.
    5. Type: Indicates the type of property being listed for rent, such as Apartment, House, Condo, etc.
    6. Price: The rental price associated with the property, indicating the cost for renting the property for a specified period.
    7. Location: The geographical location of the rental property, providing information on where the property is situated.
    8. Bedrooms: Specifies the number of bedrooms available in the rental property, indicating the sleeping accommodations.
    9. Bathrooms: Indicates the number of bathrooms available in the rental property, providing information on the sanitary facilities.
    10. Hydro: Indicators for utilities included in the rental agreement, specifying whether hydro (electricity) is included in the rental package.

11. Heat: Indicators for utilities included in the rental agreement, specifying whether heat is included in the rental package.
12. Water: Indicators for utilities included in the rental agreement, specifying whether water is included in the rental package.
13. Latitude: The geographical latitude coordinates of the rental property's location, providing precise geographic positioning.
14. Longitude: The geographical longitude coordinates of the rental property's location, providing precise geographic positioning.
15. Size: The size of the rental property, typically measured in square feet or square meters, indicating the property's spatial dimensions.
16. Agreement Type: Specifies the type of rental agreement associated with the property, such as lease, sublease, month-to-month, etc.
17. URL: The web address (URL) linking to the online listing of the rental property, providing access to additional details and images.
18. Date Posted: The date when the rental property listing was posted online, indicating when the property became available for rent.

Identified irrelevant columns for the analysis, such as 'CSDUID,' 'Title,' 'Location,' 'adId,' 'URL,' 'Date Posted,' and 'Agreement Type,' and we were dropped.

```
columns_to_drop = ['CSDUID', 'Title', 'Location', 'adId', 'URL', 'Date Posted',
'Agreement Type']
df = df.drop(columns_to_drop, axis=1)
```

Exploratory Data Analysis (EDA) Highlights:

Handling Missing Values and Data Issues:

Price Column Cleaning:

- Issue: The 'Price' column contained a combination of numerical values and "$" signs.
- Solution: Removed the "$" sign to keep only numerical values.

```
df['Price'] = df['Price'].str.replace('$','')
```

Extracting Numerical Values from Bedroom and Bathroom Columns:

- Issue: Bedroom and bathroom columns contained a combination of strings and numerical values.
- Solution: Extracted numerical values from these columns and converted them to numeric data type.

```
# Extract numeric values from 'Bedrooms' and 'Bathrooms'
data['Bedrooms'] = data['Bedrooms'].str.extract('(\d+)').astype(float)
data['Bathrooms'] = data['Bathrooms'].str.extract('(\d+\.?\d*)').astype(float)
```

Converting Data Types:

- Issue: The datatype of 'CSDNAME', 'Type', and 'Agreement Type' columns was object.
- Solution: Converted these columns to categorical datatype.

```
df['CSDNAME'] = df['CSDNAME'].astype('category')
df['Type'] = df['Type'].astype('category')
```

Handling Outliers in Size Column:

- Issue: Some properties had unrealistically small or large sizes compared to their prices.
- Solution: Removed outliers based on a threshold (e.g., below 200 sqft or above 9000 sqft).

```
df = df[(df['Size'] >= 200) & (df['Size'] <= 9000)]
```

Handling 'Not Available' Values in Size Column:

- Issue: Some entries in the 'Size' column were marked as "Not available".
- Solution: Replaced these values with NaN (Not a Number) and then filled them with the mean of the column.

```
df['Size'].replace('Not available', np.nan, inplace=True)
df['Size'].fillna(df['Size'].mean(), inplace=True)
```

Transforming Hydro, Water, and Heat Columns:

- Issue: Hydro, water, and heat attributes represented by "y" and "n".
- Solution: Converted these values into numerical equivalents (1 for "yes", 0 for "no").

```
df['Hydro'] = df['Hydro'].map({'y': 1, 'n': 0})
df['Water'] = df['Water'].map({'y': 1, 'n': 0})
df['Heat'] = df['Heat'].map({'y': 1, 'n': 0})
```

Data Visualization:

For data visualization, we utilized Looker Studio to represent the findings of our Exploratory Data Analysis (EDA) interactively and intuitively:

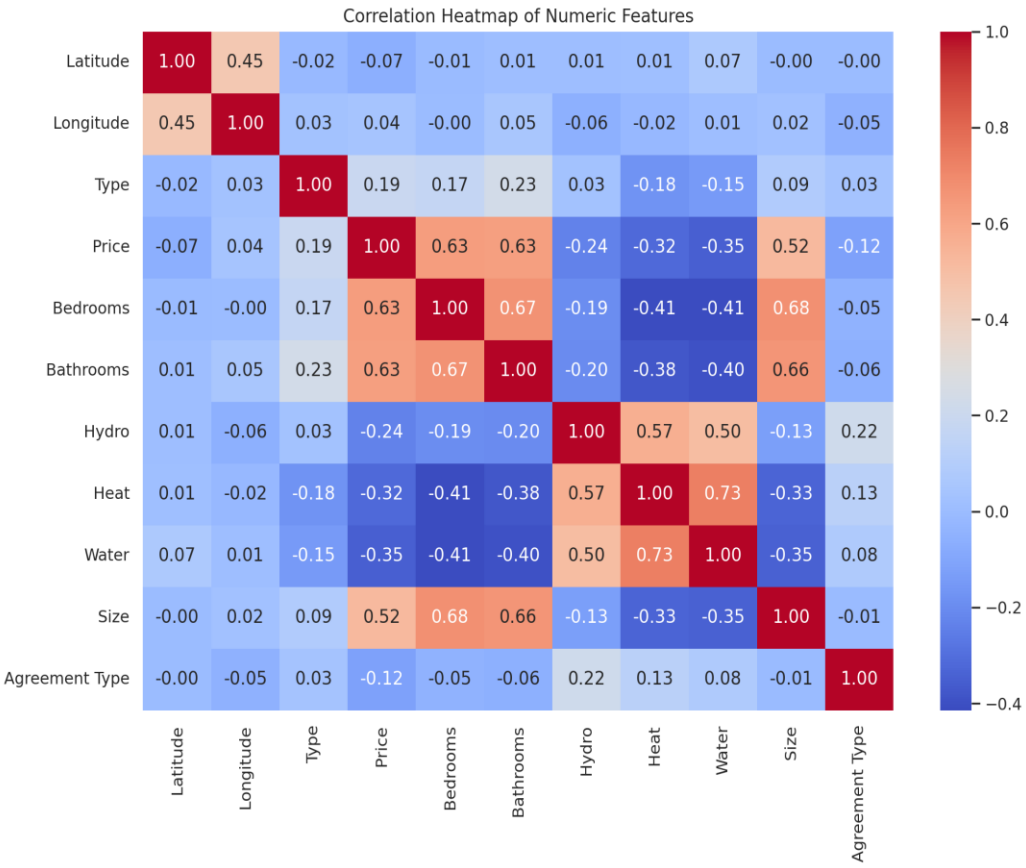
1. Bar Charts:

- We employed bar charts to illustrate categorical data, such as the distribution of property types across different regions within Ontario. This visualization allowed stakeholders to easily compare the prevalence of apartments, houses, and other property types in various areas.
- Additionally, we created bar charts to showcase the frequency of specific amenities, providing insights into renters' preferences and potential opportunities for property improvement or marketing.

2. Pie Charts:
    - Pie charts were utilized to display the composition of rental properties within each region of Ontario. By segmenting the total number of properties into categories based on property type, stakeholders could quickly grasp the relative proportions of apartments, houses, and other types of rental units.
    - This visualization helped stakeholders understand the diversity of the rental market in each region and identify potential opportunities for investment or development based on property type demand.

3. Scatter Plot:
    - We utilized scatter plots to explore the relationship between property size and rental price. By plotting each rental property as a point on the graph with its corresponding size and price, stakeholders could visually assess any patterns or trends in the data.
    - This visualization technique allowed stakeholders to identify potential correlations between property size and rental price, providing insights into pricing strategies and market dynamics.

4. Heat Map:
    - A heat map was generated to visualize the spatial distribution of rental properties across Ontario. By aggregating the density of rental listings within geographic regions, stakeholders could identify areas with high demand for rental properties.
    - This visualization enabled stakeholders to pinpoint geographic clusters of rental activity and prioritize areas for further analysis or investment based on market demand.



By employing a combination of bar charts, pie charts, scatter plots, and heat maps, we ensured that our visualizations were informative but also intuitive and actionable. Each visualization was designed to convey key insights from our EDA findings in a clear and accessible manner, empowering stakeholders to make informed decisions in the dynamic Ontario rental market.

Challenges Encountered:

Dealing with Outliers:

- Identification and Management: Recognizing outliers within the dataset, especially in numerical features like 'Size', and implementing suitable techniques to manage their impact on analysis.

Handling Missing Values:

- Understanding Missing Data: Identifying patterns and reasons behind missing values across different columns to determine appropriate imputation strategies.
- Imputation Techniques: Selecting and applying imputation methods such as mean/median imputation, forward or backward filling, or advanced techniques like multiple imputation to handle missing values effectively.

Standardizing Data Types:

- Ensuring Data Consistency: Ensuring uniformity in data types across various features to facilitate coherent analysis and modeling.

- Addressing Mixed Data Types: Tackling challenges posed by mixed data types, such as numeric and categorical values in the same column, through suitable encoding or transformation methods to maintain dataset integrity.

REFERENCES:

- Lemenkova, P. (2019). PROCESSING OCEANOGRAPHIC DATA BY PYTHON LIBRARIES NUMPY, SCIPY AND PANDAS. Aquatic Research, 73–91. https://doi.org/10.3153/ar19009
- Raschka, S., Patterson, J., & Nolet, C. (2020, April 4). Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. Information, 11(4), 193. https://doi.org/10.3390/info11040193
- Douglass, M. J. J. (2020, August 12). Book Review: Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, 2nd edition by Aurélien Géron. Physical and Engineering Sciences in Medicine, 43(3), 1135–1136. https://doi.org/10.1007/s13246-020-00913-z
- Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. Frontiers in Neuroinformatics, 8. https://doi.org/10.3389/fninf.2014.00014
- Khan, A. I., & Al-Habsi, S. (2020). Machine Learning in Computer Vision. Procedia Computer Science, 167, 1444–1451. https://doi.org/10.1016/j.procs.2020.03.355
- Waskom, M. (2021, April 6). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021
- Schillaci, M. J. (2017, November). Perfectly Python. Computing in Science & Engineering, 19(6), 51–53. https://doi.org/10.1109/mcse.2017.3971168
- Nordhausen, K. (2009, October 29). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman. International Statistical Review, 77(3), 482–482. https://doi.org/10.1111/j.1751-5823.2009.00095_18.x