IST 652 PROJECT PROPOSAL

The final project for IST652 involves locating an open data set or a group of data sets of interest, formulating an inquiry or set of inquiries that could be addressed with the data,processing the data set(s) in a Jupyter Notebook environment using Python, and conducting some analyses on the data to illuminate the inquiry. The project focuses on open data in order to ensure that your chain of transformations and analysis is reproducible.

This is the FIRST DELIVERABLE

Project Objective

Primary objectives for the project are ..

- Demonstrate your ability to write Python scripts to access and process data.
- Describe steps taken to prepare the data for analysis. For example how did you access and ingest the data, data wrangling, formatting, feature engineering and other steps.
- Develop a research questions you are hoping to answer from the data collected.
- Clearly articulate findings from analysis and summarizes impactful findings.
- Collaborate as a team.

Analysis Team

List team members below and their roles (note roles may be modified in the second deliverable)

```
--== Double-Click and Write Your Project Summary Below This Line ==--
```

Andrea Hayman and Hang Tian collaborate on this work. We will separate the work while sharing ideas upon every upcoming questions.

Codes and progress can be seen in the GitHub repo here: https://github.com/ht6631/IST_652_Project

Phase 1: Ideation

The goal of this phase is to outline the specific goals and objectives of your project; include evidence of its feasibility by including citations of resources you will use to complete the code.

Step 1: Project Summary

Write a brief summary of your project ideas, In 250 - 500 words.

`--== Double-Click and Write Below this Line ==--`
This project will be a data-oriented analysis of certain topics. There are multiple datasets on the list now and we will choose one for further analysis. According to certain background of datasets, we will ask questions and answer them as listed as in next step.

Step 2: Datasets Research

Select a dataset or a combination of datasets for your project. Many data sets are available at sites such as the World Bank (http://data.worldbank.org), the U.S. Federal Government (http://www.data.gov), - other potential sites for data sets will be provided by the instructor but it is recommended that you search for open data sets too on your own. However, do not use datasets from Kaggle.com.

Note: The number of records (rows) present in your dataset (or total combination of datasets) must exceed 4,000 with at least 8 different categories (columns) of data.

Clearly describe from where your data was located.Why is this resource an authority. Provide a shortlist of datasets your team is considering for your final project. Provide references to the dataset as applicable. Include any other components necessary.

`--== Double-Click and Write Below this Line ==--`
Among the datasets introduced in the short list below, we decide to work on the one about housing code violation in Syracuse, which is the open data from Syracuse government.

Dataset shortlist:

1. Bike-sharing dataset from UCI machine learning repository

URL: https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset
Introduction: This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information. Location: This dataset is from the UCI machine learning repository, open to all.

In [ ]:
```
# 1. Code introduction for Bike-sharing dataset
import pandas as pd
import numpy as np
bike_share=pd.read_csv('Datasets/hour.csv')
print(bike_share.shape)
print(bike_share.columns)
# There are 17379 hourly instances in the dataset with 17 columns (arrtributes) listing the hourly usage c
# See the web url for detailed information on column names and meanings.
bike_share['yr'].unique()
# There are two years' records in the dataset
```

```
(17379, 17)
Index(['instant', 'dteday', 'season', 'yr', 'mnth', 'hr', 'holiday', 'weekday',
       'workingday', 'weathersit', 'temp', 'atemp', 'hum', 'windspeed',
       'casual', 'registered', 'cnt'],
      dtype='object')
```

Out[ ]:
```
array([0, 1])
```

2. Housing code violations in Syracuse:

URL: https://data.syr.gov/datasets/2cc4e180fc6540fbb4fc6fafde311d7b_0/explore?location=43.034377%2C-76.139450%2C13.53 Introduction: This dataset contains all housing code violations reported to Code Enforcement in Syracuse with information on where, when, and what violation occurred, as well as who violated the code.

In [ ]:
```
df1 = pd.read_csv('Datasets/Code_Violations.csv')
df2 = pd.read_csv('Datasets/Code_Violations-1.csv')

code_violations = pd.concat([df1, df2], ignore_index=True)
print(code_violations.shape)
print(code_violations.columns)
```

```
# There are 98420 code violations with 24 attributes (columns) providing information on each code violatio
# The data was pulled manually, with the date range 12/26/2018 to 10/31/2023
```

```
(98420, 24)
Index(['X', 'Y', 'violation_number', 'complaint_address', 'complaint_zip',
       'SBL', 'violation', 'violation_date', 'comply_by_date',
       'status_type_name', 'complaint_number', 'complaint_type_name',
       'open_date', 'owner_name', 'inspector_id', 'Neighborhood', 'Vacant',
       'owner_address', 'owner_city', 'owner_state', 'owner_zip_code',
       'Latitude', 'Longitude', 'ObjectId'],
      dtype='object')
```

Step 2a: Objectives

What have you learned about your dataset(s) so far, and what are the questions you plan to answer with the data (a minimum of 5 questions is a good start).

```
--== Double-click and write below this line ==--
```

About the Bike-sharing dataset
This dataset records the hourly bike usage counts for two years, supplemented by weather information such as weather type of that hour, temperature and feeling temperature, relative humidity etc. Besides, there are dummy columns as indicators suggesting if that date is a holiday or on weekdays or is it a working day.

Potential questions could be:

1. What's the correlation between weather-related variables and hourly bike usage counts?
2. Is there any difference between bike usage during holiday or non-holidays?
3. If the answer to question 2 is "Yes", then what's the difference?
4. What's the hourly distribution of bike usage, is there any tedency?
5. Suggest we can get access to date-related and weather-related data, could we predict houlry bike usage for any given hour?

About the Code violations dataset
This dataset records every housing code violation in Syracuse, as well as where it occurred, when it occurred, what the violation was, as well as the name and location of who is responsible for the violation.

Potential questions could be:

1. Is there any correlation between neighborhood/location and type of violation?
2. Are the building owners generally local? Are any of them repeat offenders?
3. How did COVID impact the number of housing violations (if at all)?
4. Have the kinds of violations given out changed over time?
5. Suppose we connect our location data to demographic data, will we see a relationhship between demographic and the types/amount of violations?

References

```
--== Double-click and write below this line ==--
```