Exercise 6.1 Sourcing Data

5**. Data Source: New York City Citywide Payroll Data (Fiscal Year)**
https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e

This data provides salary information for city employees across the five boroughs of New York City.

I chose this data set for personal interest reasons and also because it fulfills the requirements of this assignment.

1. I am from NYC and so it helps me feel connected to my hometown by learning more about it through data analysis.
2. This data set has the necessary variable types to complete this project.
    - Geolocation data - in this data set location is only listed by borough. In order to convert this to latitude/longitude data, I will have to combine the Borough Boundaries dataset too.
    - Categorical data - agency name, title_description, leave_status, payroll_number
    - Continuous data - base_salary, regular_hours, regular_gross_paid, ot_hours,total_ot_paid
    - Time data - fiscal_year, agency_start_date

6. **Data Cleaning**
Based on my exploratory data analysis, I conducted the following data cleaning:

1. Continuous values and outliers:
   On intial inspection I identified negative values in pay and hours worked variables. It is highly unlikely an employee worked negative hours or earned a negative paycheck.
   The data dictionary on the data source does not explain why there are negative values in these columns.
   Total rows with negative values was approximately 1% of data.
   Final cleaning: removed the rows with negative values.
   The following chart shows min values with negative values in original data set:

```
In [7]: df.describe()
```

Out[7]:

	Base Salary	Regular Hours	Regular Gross Paid	OT Hours	Total OT Paid	Total Other Pay
count	5.109775e+06	5.109775e+06	5.109775e+06	5.109775e+06	5.109775e+06	5.109775e+06
mean	4.434074e+04	6.597381e+02	4.246115e+04	6.298963e+01	3.336576e+03	2.915953e+03
std	4.294908e+04	8.857047e+02	4.027641e+04	1.644688e+02	9.421155e+03	6.003723e+03
min	1.000000e-02	-1.872770e+03	-2.054519e+05	-2.090000e+02	-2.649388e+04	-2.815950e+05
25%	3.318000e+01	0.000000e+00	3.185280e+03	0.000000e+00	0.000000e+00	0.000000e+00
50%	4.250000e+04	0.000000e+00	3.571858e+04	0.000000e+00	0.000000e+00	3.293000e+01
75%	7.922700e+04	1.820000e+03	7.509907e+04	0.000000e+00	1.107000e+02	3.403910e+03
max	4.147070e+05	4.171430e+03	6.723089e+05	3.692900e+03	2.559999e+05	6.500000e+05

```
#output shows there are min values below zero for hours worked and amount paid. must confirm if this is bad data or if it appears regularly in data set
```

2. Duplicate rows.
   There are no duplicate rows found in this data set:

```
In [8]: df.duplicated()

Out[8]: 0          False
        1          False
        2          False
        3          False
        4          False
                   ...
        5109770    False
        5109771    False
        5109772    False
        5109773    False
        5109774    False
        Length: 5109775, dtype: bool

In [9]: df['dup_row']=df.duplicated()

In [10]: df['dup_row'].value_counts(dropna=False)

Out[10]: False    5109527
         True         248
         Name: dup_row, dtype: int64

         #output shows there are no duplicate rows
```

3. Null Values
   Nulls were found in the following columns and I cleaned as follows:
   - Payroll Number – approx 25% of data has nulls. Decided to keep this data as an "unkown" category.
   - First Name/Last Name/Mid Init – approx 12,000 rows found with incomplete name information. Removed this data from data set.
   - Work Location Borough – approx 10% of data missing this value. Decided to keep this data as "unknown" category.
   - Title Description – 55 rows found missing this value. Removed these rows from data set.
4. Consistent Values
   - Work Location Borough had mixed case values – I converted all values to all-upper-case.
5. Data Types
   - The original data contained mixed data types and would not import to a dataframe in that format. Therefore, I imported all data as strings, then converted all continuous data to float64. Python converted the date values to int64. I will keep these data types unless I find in my analysis a different data type is required.
6. Other Cleaning
   - Work Location Borough – the scope of this analysis is within the 5 boroughs of New York City. This data set contained approx 200,000 rows of data belonging to locations outside of the 5 boroughs. Examples such as Washington, DC or Albany, NY.
     Given the scope of this project, I removed these extraneous locations and I still have enough data to do my analysis.
   - Full Name/Last Name/Mid Init – given data privacy laws, I replaced these values with an Employee ID value. These values appear in the data set because, as public government employees, the public has a right to know their names, and most likely, these employees had to agree to allow their name to appear in the data. However, as the analyst for this small project, I can make an ethical decision not include these values.

To do this,

- I made assumption that the combination of 'First Name'+'Mid Init'+'Last Name'+'Agency Start Date' is representative of a unique employee.
- I created a second lookup table with each unique employee and assigned it a unique random character string.
- I merged the original data set with the new lookup table to give each row its proper Employee ID value.
- I removed the original columns with name values

7. Data Understanding

The final clean data set for this project is
4915337 rows
15 columns

The head of the data set is:

```
In [6]: df.head()
```

Out[6]:

	Fiscal Year	Payroll Number	Agency Name	Agency Start Date	Work Location Borough	Title Description	Leave Status as of June 30	Base Salary	Pay Basis	Regular Hours	Regular Gross Paid	OT Hours	Total OT Paid	Total Other Pay	employee_id
0	2020	17.0	OFFICE OF EMERGENCY MANAGEMENT	08/10/2015	BROOKLYN	EMERGENCY PREPAREDNESS MANAGER	ACTIVE	86005.0	per Annum	1820.0	84698.21	0.0	0.0	0.0	emgiazxy
1	2020	17.0	OFFICE OF EMERGENCY MANAGEMENT	09/12/2016	BROOKLYN	EMERGENCY PREPAREDNESS MANAGER	ACTIVE	86005.0	per Annum	1820.0	84698.21	0.0	0.0	0.0	qcekinkp
2	2020	17.0	OFFICE OF EMERGENCY MANAGEMENT	02/22/2016	BROOKLYN	EMERGENCY PREPAREDNESS MANAGER	ACTIVE	86005.0	per Annum	1820.0	84698.21	0.0	0.0	0.0	vihcvktl
3	2020	17.0	OFFICE OF EMERGENCY MANAGEMENT	09/16/2013	BROOKLYN	EMERGENCY PREPAREDNESS MANAGER	ACTIVE	86005.0	per Annum	1820.0	84698.21	0.0	0.0	0.0	ototwwxb
4	2020	17.0	OFFICE OF EMERGENCY MANAGEMENT	04/30/2018	BROOKLYN	EMERGENCY PREPAREDNESS MANAGER	ACTIVE	86005.0	per Annum	1820.0	84698.21	0.0	0.0	0.0	oskpiihm

Exploratory descriptive statistics show no outliers or negative values:

```
In [8]: df.describe()
```

Out[8]:

	Fiscal Year	Payroll Number	Base Salary	Regular Hours	Regular Gross Paid	OT Hours	Total OT Paid	Total Other Pay
count	4.915337e+06	3.234063e+06	4.915337e+06	4.915337e+06	4.915337e+06	4.915337e+06	4.915337e+06	4.915337e+06
mean	2.018074e+03	5.736013e+02	4.231114e+04	6.705228e+02	4.166508e+04	6.448972e+01	3.409022e+03	2.954736e+03
std	2.577118e+00	3.008655e+02	4.194309e+04	8.886410e+02	3.953947e+04	1.663559e+02	9.521246e+03	5.727339e+03
min	2.014000e+03	2.000000e+00	1.000000e-02	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	2.016000e+03	3.000000e+02	3.318000e+01	0.000000e+00	3.131120e+03	0.000000e+00	0.000000e+00	0.000000e+00
50%	2.018000e+03	7.420000e+02	4.100300e+04	0.000000e+00	3.518840e+04	0.000000e+00	0.000000e+00	4.311000e+01
75%	2.020000e+03	7.470000e+02	7.656500e+04	1.820000e+03	7.363837e+04	1.500000e+00	1.496900e+02	3.389850e+03
max	2.022000e+03	9.960000e+02	4.147070e+05	4.171430e+03	6.723089e+05	3.692900e+03	2.559999e+05	6.500000e+05

The info shows proper data types:

```
In [9]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4915337 entries, 0 to 4915336
Data columns (total 15 columns):
 #   Column                     Dtype
---  ------                     -----
 0   Fiscal Year                int64
 1   Payroll Number             float64
 2   Agency Name                object
 3   Agency Start Date          object
 4   Work Location Borough      object
 5   Title Description          object
 6   Leave Status as of June 30 object
 7   Base Salary                float64
 8   Pay Basis                  object
 9   Regular Hours              float64
 10  Regular Gross Paid         float64
 11  OT Hours                   float64
 12  Total OT Paid              float64
 13  Total Other Pay            float64
 14  employee_id                object
dtypes: float64(7), int64(1), object(7)
memory usage: 562.5+ MB
```

There are no duplicates:

```
In [10]: df.duplicated()
```

```
Out[10]: 0          False
         1          False
         2          False
         3          False
         4          False
                    ...
         4915332    False
         4915333    False
         4915334    False
         4915335    False
         4915336    False
         Length: 4915337, dtype: bool
```

There are no null values:

```
In [14]: df.isnull()
```

Out[14]:

	Fiscal Year	Payroll Number	Agency Name	Agency Start Date	Work Location Borough	Title Description	Leave Status as of June 30	Base Salary	Pay Basis	Regular Hours	Regular Gross Paid	OT Hours	Total OT Paid	Total Other Pay	employee_id
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4915332	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4915333	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4915334	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4915335	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4915336	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False

4915337 rows × 15 columns

Description of each variable

Variable Name	Example Value	Description
Fiscal Year	2020	Year in which payroll info was recorded.
Agency Name	OFFICE OF EMERGENCY MANAGEMENT	Department where employed.
Agency Start Date	08/10/2015	Date on which employee began, uses mm/dd/yyyy format.
Work Location Borough	BROOKLYN	NYC Borough where employed.
Title Description	EMERGENCY PREPAREDNESS MANAGER	Title of employment.
Leave Status as of June 30	ACTIVE	Category of current employment status.
Base Salary	86000.50	Dollar amount of Base Salary (may not be amount actually earned).
Pay Basis	per Annum	Category how pay is distributed.
Regular Hours	1820.0	Number of hours worked per year to earn the regular salary.
Regular Gross Paid	84698.21	Dollar amount actually paid for regular working hours.
OT Hours	0.0	Number hours worked above regular working hours and which earn extra pay.
Total OT Paid	0.0	Dollar amount paid for overtime hours.
Total Other Pay	0.0	Dollar amount paid for any other reason.
employee_id	hoiwnsit	8 character random string unique identifier for employee.

8. Data Ethics

The original data set included full name information for each employee.

Data ethics include protecting the privacy of individuals included in the data set.

While it is legal public information to include the full name of government employees in the data, I as the analyst for this project can choose to omit this information and protect the privacy of these people.

To eliminate private information, I replaced the name data with a random unique identifier so that names will not appear in my analysis.

9. Possible Analysis Questions

Given the variables in this data set,

- Which titles or agencies pay the most or the least?
- Do employees receive the same pay in all 5 boroughs?
- Do employees earn more money based on how long they stay at the job?
- How often do employees change title or agency?
- Has the popularity of certain jobs changed over time?
- What is the location of each job title?
- Does one borough hire more of less of a certain job title than the other boroughs?