Analysis Notebook

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn as sk
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

```python
from ucimlrepo import fetch_ucirepo

# fetch dataset
seoul_bike_sharing_demand = fetch_ucirepo(id=560)

# data (as pandas dataframes)
X = seoul_bike_sharing_demand.data.features
y = seoul_bike_sharing_demand.data.targets

# metadata
print(seoul_bike_sharing_demand.metadata)

# variable information
print(seoul_bike_sharing_demand.variables)
```

```
{'uci_id': 560, 'name': 'Seoul Bike Sharing Demand', 'repository_url': 'https://archive.ics.
                     name     role         type demographic description  \
0                    Date  Feature         Date        None        None
1       Rented Bike Count  Feature      Integer        None        None
2                    Hour  Feature      Integer        None        None
3             Temperature  Feature   Continuous        None        None
4                Humidity  Feature      Integer        None        None
5              Wind speed  Feature   Continuous        None        None
6              Visibility  Feature      Integer        None        None
7   Dew point temperature  Feature   Continuous        None        None
8         Solar Radiation  Feature   Continuous        None        None
9                Rainfall  Feature      Integer        None        None
10               Snowfall  Feature      Integer        None        None
```

```
11             Seasons  Feature  Categorical        None        None
12             Holiday  Feature       Binary        None        None
13     Functioning Day   Target       Binary        None        None

    units missing_values
0    None             no
1    None             no
2    None             no
3       C             no
4       %             no
5     m/s             no
6     10m             no
7       C             no
8   Mj/m2             no
9      mm             no
10     cm             no
11   None             no
12   None             no
13   None             no
```

Instead of using the functional days as a target, we will use it as a predictor and try to predict the rented bike count to forecast demand for bikes. Let's do some data cleaning and preprocessing to make these new matrices as well as finalize our features for model fitting.

```
target = X['Rented Bike Count']
X = X.drop(columns=['Rented Bike Count'])
X['Functional Day'] = y.astype('category')
y = target

X['Date'] = pd.to_datetime(X['Date'], format='%d/%m/%Y')
X['Day'] = X['Date'].dt.day
X['Month'] = X['Date'].dt.month
X['Year'] = X['Date'].dt.year
X = X.drop(columns=["Date"])

X = pd.get_dummies(X, drop_first=True)
X = X.astype('float')
```

```
X.head()
```

	Hour	Temperature	Humidity	Wind speed	Visibility	Dew point temperature	Solar Radiation
0	0.0	-5.2	37.0	2.2	2000.0	-17.6	0.0
1	1.0	-5.5	38.0	0.8	2000.0	-17.6	0.0
2	2.0	-6.0	39.0	1.0	2000.0	-17.7	0.0
3	3.0	-6.2	40.0	0.9	2000.0	-17.6	0.0

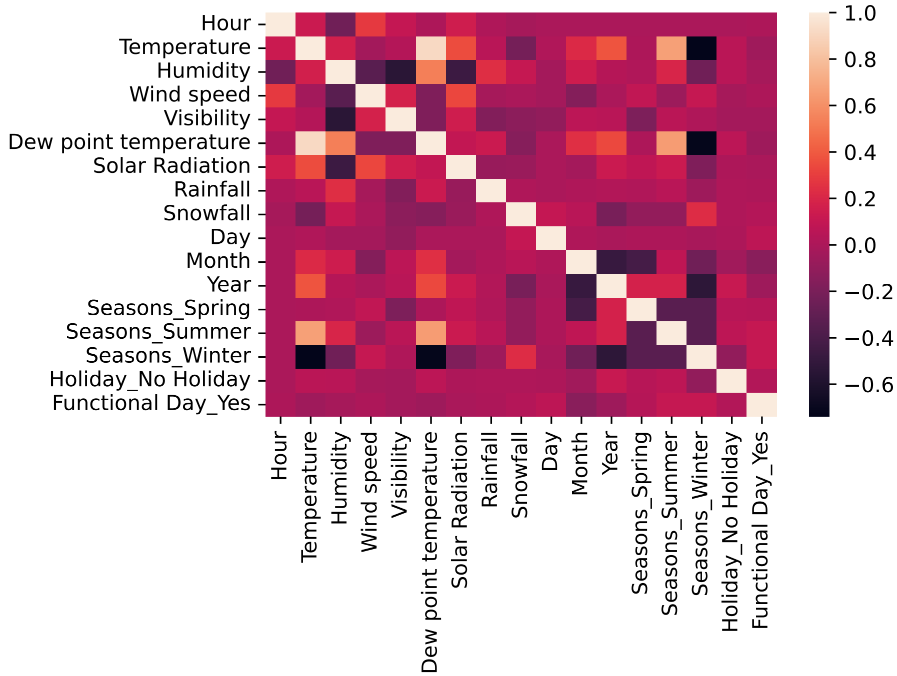
	Hour	Temperature	Humidity	Wind speed	Visibility	Dew point temperature	Solar Radiation
4	4.0	-6.0	36.0	2.3	2000.0	-18.6	0.0

```
y.head()
```

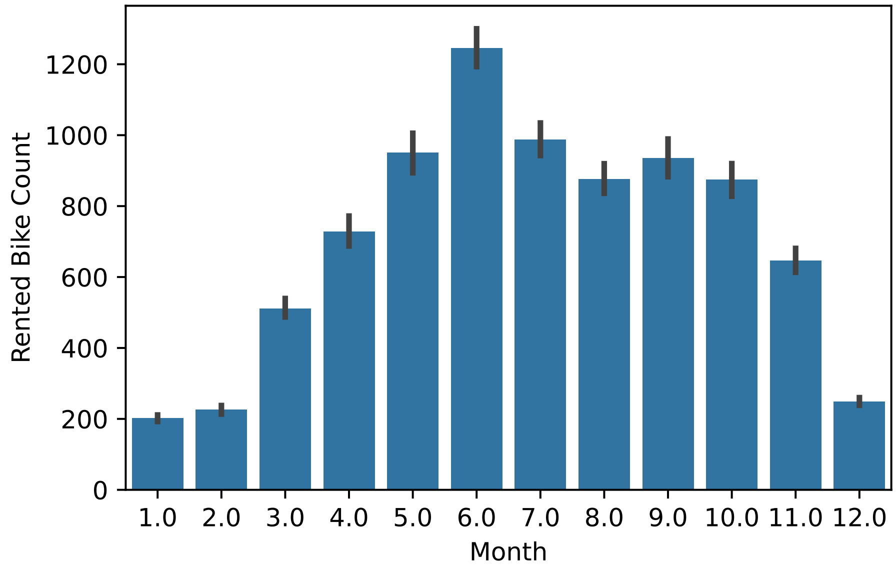
```
0    254
1    204
2    173
3    107
4     78
Name: Rented Bike Count, dtype: int64
```

Now let's do some visualization with the dataset just to see if there are any useful patterns.

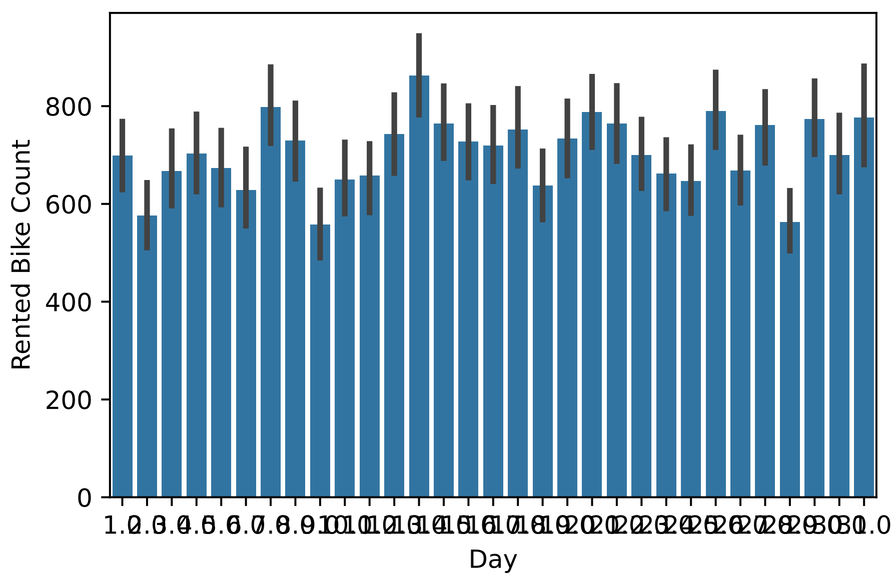
```
sns.heatmap(X.corr())
```



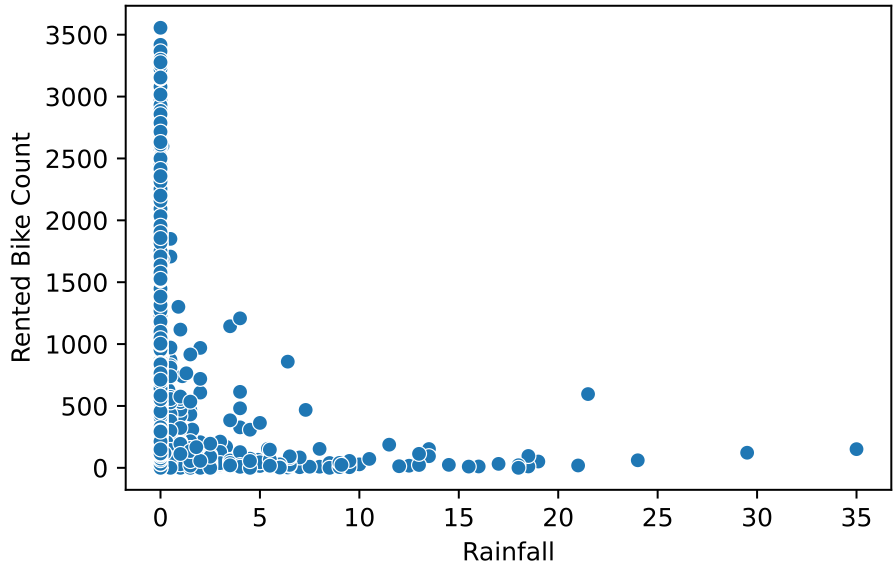

```
sns.barplot(x='Month', y=y, data=X)
```



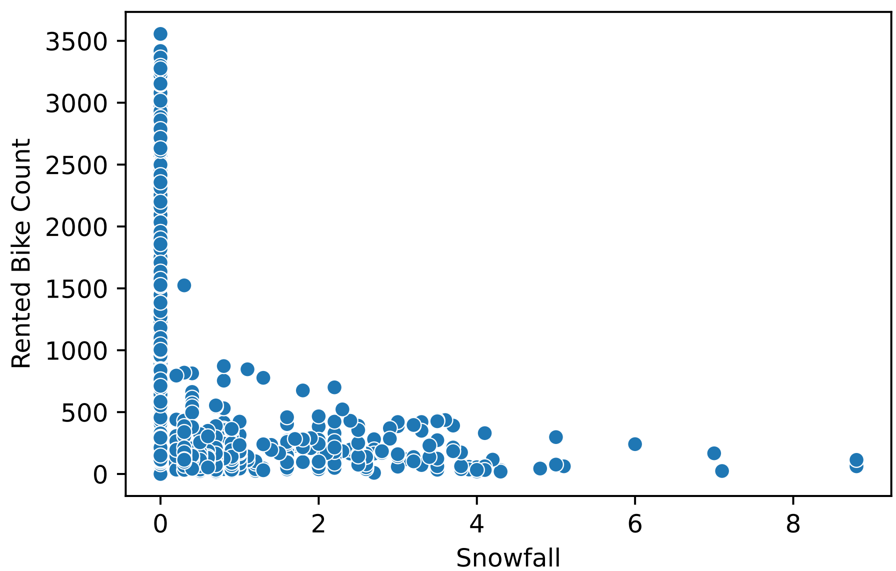

```
sns.barplot(x="Day", y=y, data=X)
```



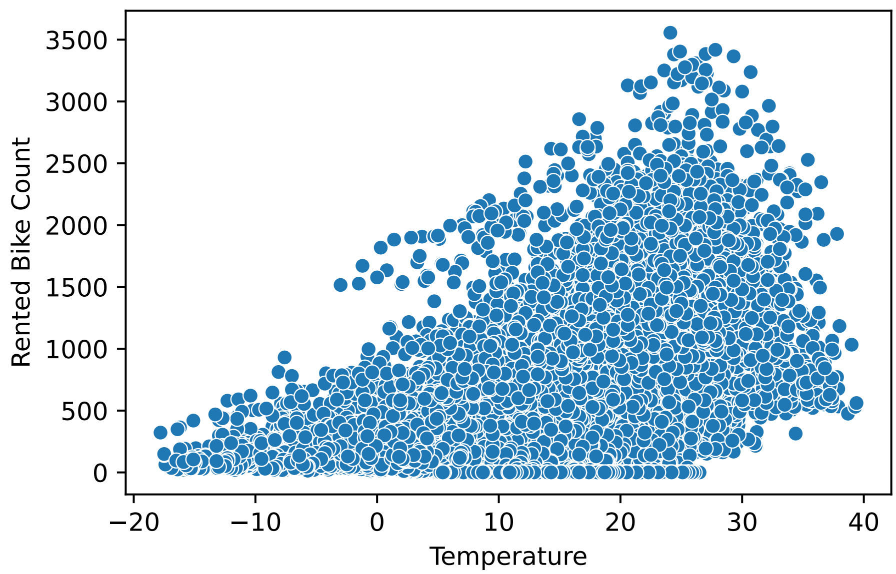

```
sns.scatterplot(x='Rainfall', y=y, data=X)
```



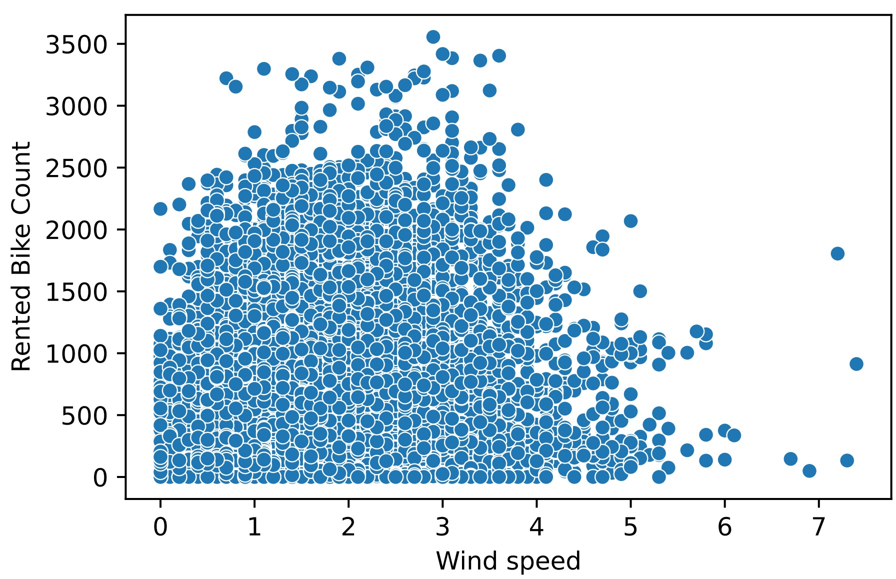

```
sns.scatterplot(x="Snowfall", y=y, data=X)
```




```
sns.scatterplot(x="Temperature", y=y, data=X)
```




```python
sns.scatterplot(x="Wind speed", y=y, data=X)
```




```python
lm = sm.OLS(y, X).fit()

print(lm.summary())
```

```
                            OLS Regression Results
```

```
=======================================================================================
Dep. Variable:      Rented Bike Count   R-squared (uncentered):                   0.795
Model:                            OLS   Adj. R-squared (uncentered):              0.795
Method:                 Least Squares   F-statistic:                              1998.
Date:                Fri, 10 Jan 2025   Prob (F-statistic):                        0.00
Time:                        00:02:33   Log-Likelihood:                         -65594.
No. Observations:                8760   AIC:                                 1.312e+05
Df Residuals:                    8743   BIC:                                 1.313e+05
Df Model:                          17
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Hour                     27.4627      0.735     37.376      0.000      26.022      28.903
Temperature              15.9652      3.662      4.359      0.000       8.786      23.144
Humidity                -10.9818      1.031    -10.651      0.000     -13.003      -8.961
Wind speed               19.5192      5.100      3.827      0.000       9.522      29.517
Visibility                0.0070      0.010      0.701      0.483      -0.013       0.027
Dew point temperature    11.2057      3.834      2.923      0.003       3.691      18.721
Solar Radiation         -78.2988      7.610    -10.289      0.000     -93.216     -63.381
Rainfall                -58.3407      4.269    -13.666      0.000     -66.709     -49.973
Snowfall                 32.8892     11.332      2.902      0.004      10.677      55.102
Day                      -1.0765      0.539     -1.998      0.046      -2.132      -0.021
Month                     4.0784      1.813      2.249      0.025       0.524       7.633
Year                     -0.0468      0.050     -0.939      0.348      -0.145       0.051
Seasons_Spring         -112.1860     17.465     -6.423      0.000    -146.422     -77.950
Seasons_Summer         -140.8455     18.260     -7.713      0.000    -176.640    -105.051
Seasons_Winter         -349.9548     21.221    -16.491      0.000    -391.553    -308.357
Holiday_No Holiday      120.4375     21.638      5.566      0.000      78.023     162.852
Functional Day_Yes      935.7315     26.710     35.034      0.000     883.374     988.089
==============================================================================
Omnibus:                     1426.583   Durbin-Watson:                   0.510
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2863.536
Skew:                           0.990   Prob(JB):                         0.00
Kurtosis:                       4.981   Cond. No.                     1.53e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constan
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 1.53e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

With the p values produced and the large condition number, it is clear we need to perform some form of feature selection. I will use lasso regression to see which variables are most important in the prediction.

```
lasso = sm.OLS(y, X).fit_regularized(method='elastic_net', alpha=0.1, L1_wt=1)

print("Lasso Regression Coefficients:")
print(lasso.params)
```

```
Lasso Regression Coefficients:
Hour                         26.240363
Temperature                   9.308941
Humidity                    -11.552014
Wind speed                   22.244678
Visibility                    0.031987
Dew point temperature        16.886413
Solar Radiation             -62.493750
Rainfall                    -57.592126
Snowfall                     27.416864
Day                          -0.000368
Month                         7.951308
Year                          0.035022
Seasons_Spring              -60.812330
Seasons_Summer              -93.293391
Seasons_Winter             -309.429074
Holiday_No Holiday          140.319947
Functional Day_Yes          734.396895
dtype: float64
```

```
selected = lasso.params[lasso.params > 0.05]
print(selected)
```

```
Hour                         26.240363
Temperature                   9.308941
Wind speed                   22.244678
Dew point temperature        16.886413
Snowfall                     27.416864
Month                         7.951308
Holiday_No Holiday          140.319947
Functional Day_Yes          734.396895
dtype: float64
```

```
X = X[selected.index]

smaller_lm = sm.OLS(y, X).fit()

print(smaller_lm.summary())
```

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:      Rented Bike Count   R-squared (uncentered):                   0.748
```

```
Model:                            OLS   Adj. R-squared (uncentered):              0.748
Method:                 Least Squares   F-statistic:                              3251.
Date:                Fri, 10 Jan 2025   Prob (F-statistic):                        0.00
Time:                        00:02:34   Log-Likelihood:                         -66499.
No. Observations:                8760   AIC:                                  1.330e+05
Df Residuals:                    8752   BIC:                                  1.331e+05
Df Model:                           8
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Hour                      25.2777      0.792     31.927      0.000      23.726      26.830
Temperature               45.1963      1.145     39.459      0.000      42.951      47.442
Wind speed               -30.3990      5.376     -5.654      0.000     -40.938     -19.860
Dew point temperature    -19.4852      1.044    -18.662      0.000     -21.532     -17.439
Snowfall                 -27.4366     12.194     -2.250      0.024     -51.341      -3.533
Month                     -1.7357      1.453     -1.195      0.232      -4.584       1.113
Holiday_No Holiday      -244.0583     19.780    -12.338      0.000    -282.833    -205.284
Functional Day_Yes       236.0189     20.421     11.558      0.000     195.989     276.049
==============================================================================
Omnibus:                     1326.519   Durbin-Watson:                   0.390
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2508.533
Skew:                           0.952   Prob(JB):                         0.00
Kurtosis:                       4.802   Cond. No.                         121.
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constar
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Based on the above results, the month predictor is no longer significant so we remove that as well. This gives us the final linear model below.

```python
X = X.drop(columns=['Month'])

final_lm = sm.OLS(y, X).fit()

print(final_lm.summary())
```

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:      Rented Bike Count   R-squared (uncentered):                   0.748
Model:                            OLS   Adj. R-squared (uncentered):              0.748
Method:                 Least Squares   F-statistic:                              3715.
Date:                Fri, 10 Jan 2025   Prob (F-statistic):                        0.00
Time:                        00:02:34   Log-Likelihood:                         -66500.
No. Observations:                8760   AIC:                                  1.330e+05
```

```
Df Residuals:                    8753   BIC:                         1.331e+05
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Hour                       25.2127      0.790     31.919      0.000      23.664      26.761
Temperature                45.0539      1.139     39.548      0.000      42.821      47.287
Wind speed                -29.9506      5.363     -5.584      0.000     -40.464     -19.438
Dew point temperature     -19.4854      1.044    -18.662      0.000     -21.532     -17.439
Snowfall                  -29.0629     12.118     -2.398      0.016     -52.818      -5.308
Holiday_No Holiday       -248.5649     19.418    -12.801      0.000    -286.628    -210.501
Functional Day_Yes        231.0510     19.993     11.556      0.000     191.859     270.243
==============================================================================
Omnibus:                     1326.920   Durbin-Watson:                   0.390
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2516.414
Skew:                           0.951   Prob(JB):                         0.00
Kurtosis:                       4.810   Cond. No.                         117.
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constan
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```