

Fall 2018

## Jupyter: Intro to Data Science - Assignment 5, Part 2 Linear Regression Practice

Grant Long  
*CUNY City College*

NYC Tech-in-Residence Corps

Follow this and additional works at: [https://academicworks.cuny.edu/cc\\_oers](https://academicworks.cuny.edu/cc_oers)



Part of the [Computer Sciences Commons](#)

**[How does access to this work benefit you? Let us know!](#)**

---

### Recommended Citation

Long, Grant and NYC Tech-in-Residence Corps, "Jupyter: Intro to Data Science - Assignment 5, Part 2 Linear Regression Practice" (2018). *CUNY Academic Works*.  
[https://academicworks.cuny.edu/cc\\_oers/164](https://academicworks.cuny.edu/cc_oers/164)

This Assignment is brought to you for free and open access by the City College of New York at CUNY Academic Works. It has been accepted for inclusion in Open Educational Resources by an authorized administrator of CUNY Academic Works. For more information, please contact [AcademicWorks@cuny.edu](mailto:AcademicWorks@cuny.edu).

# Assignment 5, Part 2: Linear Regression Practice

Assignment 5 builds on the StreetEasy dataset we discussed in class, and looks at a few other relationships in the data.

```
In [ ]: import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm

%matplotlib inline
```

```
In [ ]: se_df = pd.read_csv('https://grantmlong.com/data/streeteasy_rents_june2016.
se_df.head()
```

```
In [ ]:
```

## Question 1: Exploring Bedrooms and Bathrooms

- Create **two** separate scatterplots showing the relations between rents and bedrooms, and rents and bathrooms.
- In two to three sentences, explain any problem this might pose for predicting rents.

```
In [ ]: # your code for scatterplots here
```

--- your written response here ---

```
In [ ]:
```

## Question 2: Running a regression

- Using `statsmodels` (and the code provided below), fit a regression with a **constant**, **bedrooms**, **bathrooms**, and **time to subway** as independent variables and **rent** as the dependent variable.
- Print the regression results.

```
In [ ]: # Add a constant to our existing dataframe for modeling purposes
se_df = sm.add_constant(se_df)

est = sm.OLS(se_df['rent'],
             se_df[['XXXX', 'XXXX', 'XXXX', 'XXXX']]
             ).fit()

print(XXXX)
```

In [ ]:

### Question 3: Interpreting Results

- Write three sentences identifying the coefficient for each independent variable and explaining how each of the three predictors relates to rent.
- Briefly explain whether you believe each of these predictors are statistically significant and why.

--- your written response here ---

In [ ]:

In [ ]:

### Question 4: Identifying Pitfalls

- Create a scatterplot of the values for bedrooms with the values for bathrooms.
- Identify which of the required assumptions for linear regression that we discussed in class is challenged by this scatterplot.

In [ ]:

```
# your code for scatterplots here
```

--- your written response here ---

In [ ]:

In [ ]:

### Question 5: Overall Fit

- From the regression output produced in question 2, which of the metrics provided in the output illustrates the *overall fit* of regression?
- What is the value of the metric?
- Briefly explain whether you think this regression a formulated above is a good predictor of rent.

--- your written response here ---

--- your written response here ---

--- your written response here ---

In [ ]: