Coursera Intro to Data Science - Final Assignment

Analyzing the NYC Subway Dataset

Questions

- Note: I think maybe the course content changed since these questions were publish, or perhaps the questions apply to the paid edition of the course. They seem to assume the completion of a stand-alone project, whereas the current, free edition of the course has only "bite-sized" labs. Some of these questions don't seem to fit very well with the bite-sized labs.

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

- "Problem Set 3: Analyzing Subway Data > 3 - Mann-Whitney U-Test" used the Mann-Whitney U Test.
- Mann-Whitney U does a one-sided P-value comparison..
- Null hypothesis was that the means did not differ statistically.
- P-critical was 0.05.

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

- This did not assume a normal distribution of data.

1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

- Rain_mean, Dry_Mean, U, p = (1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721)

1.4 What is the significance and interpretation of these results?

- The means differ by about 15 riders, and the p value indicates there is only a 2% chance that this could happen by random chance, if the means really were the same.

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

- Gradient Descent, using code I wrote for an earlier lab.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

- rain, precipi, Hour, meantempi, (and the units)
  - It looks like the results would have been much better if I had filtered out the units and used just rain, precipi, Hour, meantempi.
- No dummy variables (although using them might have helped).

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.

- I used these features because they were provided by the framework of the bite-sized assignments. Because they were provided, I took the path of least resistance.

2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?

```
[  2.92398062e+00   1.46526720e+01   4.67708502e+02
-6.22179395e+01
```

```
   1.50048977e+02  -1.51041835e+01  -2.78178403e+01
-1.78111613e+01
  -3.82710650e+00  -1.51064914e+01  -3.03405205e+01
-2.81397316e+01
  -2.72977045e+01   1.66072856e+02   2.13089119e+02
 2.67066331e+02
   5.16152726e+01   8.82106563e+01   1.45131256e+01
-6.62061400e+00
   1.50151940e+02   1.81381142e+02   9.12824534e+01
 1.66301402e+02
   1.95134027e+02   3.31670190e+02   2.52955523e+02
 6.95647250e+01
   1.23023878e+02   4.96124717e+01   6.63699366e+01
 2.71275019e+02
   3.91722520e+01   9.66980021e+01   1.13481504e+02
 2.03923524e+02
   6.25995495e-01   1.05820526e+02  -1.37963537e+01
-1.80013053e+00
  -3.27954482e+01  -1.57338803e+01   2.40551279e+01
 9.27339967e+01
  -5.49078674e+00  -3.80039668e+01   1.01906404e+02
 4.57223135e+01
   2.48437131e+02   2.31882935e+02   2.83139436e+01
 7.94475649e+01
   1.31061801e+02   1.31300761e+02   4.18095222e+00
 7.96368978e+01
   2.29892511e+00   2.06037678e+02   2.14411380e+01
 1.19340398e+02
  -1.16613447e+01   1.20077190e+01  -6.26584061e+00
-1.54628210e+01
   1.10967381e+02  -8.84860608e-01  -1.17517123e+01
-6.88583093e+00
  -2.55555792e+01  -1.11363180e+00  -1.80672170e+01
-1.29962705e+01
   1.48668764e+01   7.63598907e+01   1.31615020e+02
 8.81024335e+01
   7.10038324e+00   8.62353968e+01   2.55269667e+02
 2.70952300e+01
   4.57764855e+01   5.51407232e+00  -1.77230574e+01
-1.96240474e+01
  -2.49486008e+01   1.09359230e+01   2.48741776e+01
 1.92440070e+01
   1.94854288e+01   2.66630818e+01   4.30533932e+01
 6.94323103e+01
   5.00310592e+01   4.39223928e+01  -1.76023459e+01
 7.04055675e+01
   1.38167300e+02   1.06162644e+01   3.59649577e+01
 7.88106275e+01
  -3.53738569e-01  -3.29696161e+00   1.55927816e+02
 1.69053707e+01
   7.18854459e+01   5.17933579e+01   2.25060137e+01
```

```
 6.21093692e+01
  -3.54641138e+00   1.05705688e+00   8.42167769e+01
-3.86817297e+00
   1.22408491e+01   2.29820639e+01   1.66715245e+01
 1.60787783e+01
   5.10221678e+01   5.38004964e+01  -1.49439231e+01
-1.16135349e+01
   2.89759258e+01   1.29025343e+02  -2.06271758e+01
 1.31320798e+01
  -1.65639606e+01   9.10323193e+01   8.02906034e+01
 1.11730913e+01
  -1.92383723e+01   7.82542916e+00   2.99968282e+01
 5.03477387e+01
   1.65183577e+02   1.65060583e+01   1.17815897e+01
 1.01895050e+02
   7.44241689e+01   4.91928463e+01   8.38284634e+01
-1.83678785e+01
   1.55278282e+01   4.21211046e+01  -1.03792582e+01
 2.62002267e+00
  -2.80226986e+00   3.33085928e+01   2.09961672e+01
 5.03164568e+01
   1.31932560e+00  -1.36045026e-01   3.62466188e-02
 1.11634842e+01
   1.22038517e+02   1.28670551e+01   6.66731439e+01
 1.33745672e+01
   9.60670502e+01   9.01344169e+01  -4.85764483e+01
 5.62686462e+01
   6.26727112e+01   1.48694711e+02   3.35068505e+01
 5.00414940e+02
   8.68066204e+00  -2.10257955e+00  -2.56847487e+00
 1.64480655e+01
   1.44297465e+02   6.63098422e+01   1.09567641e+02
 1.91881373e+02
   2.78125290e+02   3.83518594e+01   2.03687137e+01
 5.09909874e+01
  -5.78887965e-01  -5.46397265e+00   5.06683758e+00
 7.42743335e+00
   3.16116046e+01   5.22556678e+01   1.97528601e+01
 4.66587287e+01
   3.24473270e+01   3.13032551e+01   1.63326803e+01
 2.90693854e+01
   1.60710579e+02   6.53772896e+00   2.06010322e+01
 5.48789870e+01
  -1.03006272e+01   3.52467369e+01   7.09050598e+01
 5.47745857e+01
   9.79737628e+00   2.67633781e+01   2.66819218e+01
 2.87386871e+01
   3.99245233e+01   7.37477117e+01  -3.00408986e+00
-1.44469574e+01
   2.79650790e+01   2.95782867e+01   1.19019670e+01
```

```
-1.08724998e+01
   1.60719642e+01  -1.33071187e+01   2.40376541e+01
 1.71926617e+01
  -1.13588499e+01  -3.24544547e+00   6.72658276e+00
 7.06604331e+01
   4.23561090e+01  -8.01175225e+00  -1.49990063e+01
-8.78100551e+00
  -2.37672549e+00   6.72927256e+00  -1.51755343e+01
-1.75937192e+01
  -3.64694329e+00   1.04490832e+01  -3.06820340e+00
-2.13828360e+01
   1.15044764e+02   2.90499794e+01  -1.28082169e+01
 4.80749249e+01
  -6.43462805e+00   1.02582351e+02  -2.50690566e+01
-1.63062202e+01
   2.86130420e+01   1.69578970e+01  -7.18142286e+00
-2.52714068e+01
   8.31163186e+01   2.86389022e+01  -3.35894490e-01
 1.12344963e+01
  -6.46018722e+00  -1.74766967e+01   6.13660826e+01
-8.63953967e+00
   1.72718038e+00   3.98900587e+01   3.66611207e+01
-6.17733830e+00
   1.77550634e+01   1.04675884e+01  -7.18810321e+00
-3.56617357e+01
  -2.24938334e+01  -8.52652410e+00   7.59911032e-01
 4.62456576e+00
   5.57623298e+01  -3.87211542e+00  -1.71708270e+01
-1.89741611e+01
   3.16461361e+01   1.14515983e+01  -5.26717163e+00
-7.39687157e+00
   1.84135764e+01  -5.93125845e+00  -2.45820155e+01
-1.57118031e+01
  -1.11708169e+01   1.64393144e+01   1.73837628e+01
-1.10113821e+01
  -5.21955466e+00  -1.81140993e+01  -2.26540482e+01
-2.76647704e+00
   3.56673714e+01  -1.60159757e+01   2.98461187e+01
 2.29884191e+01
  -2.37929126e+01   2.45429952e+02  -2.57666107e+00
-1.55981971e+01
  -1.85741875e+01  -9.11463740e+00  -9.77392606e+00
-2.59040332e+01
   6.47952316e+01   3.17516647e+00   4.49564729e+01
 2.27908025e+01
  -2.30662784e+00  -3.60378307e+01  -1.79093664e+01
-3.70940451e+00
  -4.22550003e+00   1.97467272e+01  -1.81003256e+01
-2.67692378e+01
  -2.52117889e+01  -2.09685883e+01  -1.53484328e+01
```

```
-2.95207357e+01
  -1.50109966e+01  -1.77590991e+01   4.19232583e+01
-1.48460812e+01
  -5.33563180e+00   2.89895748e+01   1.15778138e+01
 2.60615941e+01
  -2.54795330e+01  -2.14884481e+01  -3.22301004e+00
-2.55735966e+01
  -2.68592085e+01   2.50492296e+00  -3.14189352e+01
-1.38648866e+01
  -1.17163150e+01   2.38271400e-01  -1.69316116e+01
-2.65934826e+01
  -3.30000680e+01  -4.09388736e+01  -2.37005601e+01
-2.66587511e+01
  -2.52111821e+01  -1.89995888e+01  -7.43492697e+00
-1.77254514e+01
  -1.84646960e+01   6.88263540e-02  -6.43422348e+00
-3.02489437e+01
  -8.63456850e+00  -2.65950082e+01  -2.45560331e+01
-1.76573585e+01
  -2.75102374e+01  -3.20560426e+01  -7.38477272e+00
-3.52012537e+01
  -2.58174990e+01   5.57228671e+01  -3.13940428e+01
-3.96081119e+00
  -9.82232561e+00  -2.06761669e+01  -8.64652986e+00
-1.93162138e+01
  -1.98290293e+01  -2.22590197e+01  -1.76862690e+01
-1.45367911e+01
  -1.59402972e+01  -1.14422889e+01  -1.46165877e+01
-1.22320571e+01
  -8.57764806e+00  -2.04223404e+01  -2.22987259e+01
 7.41479002e+00
   8.21672738e+00  -2.02835359e+01  -6.57019285e+00
-7.64812931e+00
  -4.26073602e-01  -2.60150458e+01  -2.77392980e+01
-2.74593788e+00
  -2.06771622e+00  -1.43276684e+00  -4.76246626e+00
-2.72040690e+01
   2.46829569e+01   5.43973047e+00  -4.72477894e+00
-1.57853028e+01
   1.18737442e-01  -1.25911777e+01  -1.91415337e+01
-2.87506199e+01
  -1.19308332e+01  -2.52758365e+01  -3.29973179e+01
-2.56932053e+01
  -2.69425117e+01  -2.31969460e+01  -1.04640497e+01
-1.63455270e+01
   1.10341777e+00  -3.50342263e+01   7.72466058e+00
-1.77872512e+01
  -2.27666219e+01  -2.87829997e+01  -2.53520955e+01
-3.40200700e+01
  -4.67921051e+01  -3.00631021e+01  -3.03755890e+01
```

```
-3.66659085e+01
  -3.94676403e+01  -1.85643229e+01  -2.41963528e+01
-3.24984311e+01
  -2.33215131e+01  -2.57757079e+01  -2.33830335e+01
-3.70482702e+01
  -3.04501205e+01  -3.12934992e+01  -1.19141325e+01
-1.54377113e+01
  -1.80132550e+01  -2.26994061e+01  -2.61288907e+01
-1.95216924e+01
  -2.86648011e+01  -3.15231402e+01  -1.69257037e+01
-2.03155413e+01
  -1.55249279e+01  -2.39118819e+01  -2.16670019e+01
-2.37745319e+01
  -4.27971316e+00  -2.63205000e+01  -3.84155187e+00
-1.36450122e+01
  -2.01281768e+01  -4.14871372e+01  -1.43512544e+01
-2.85127638e+01
  -7.33504374e+00   1.69717762e+02   1.91311452e+01
-2.94194679e+01
  -2.50590366e+01  -2.79625975e+01  -2.61591921e+01
 1.43779868e+01
   1.03360548e+02   3.41207957e+01   6.78632689e+01
-3.13487939e+01
  -2.14761798e+01  -2.15114698e+01  -2.64818431e+01
-2.47460624e+01
  -1.62082154e+02  -2.06914281e+02  -1.08074254e+02
-1.78889027e+02
  -1.33497294e+02  -1.25293661e+02  -1.07118617e+02
-7.86617495e+01
  -6.48198796e+01  -3.18393401e+02  -2.41395539e+02
-1.50927305e+02
  -1.51524518e+02   1.10060866e+03]
```

2.5 What is your model's R^2 (coefficients of determination) value?

- 0.47924770782

2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

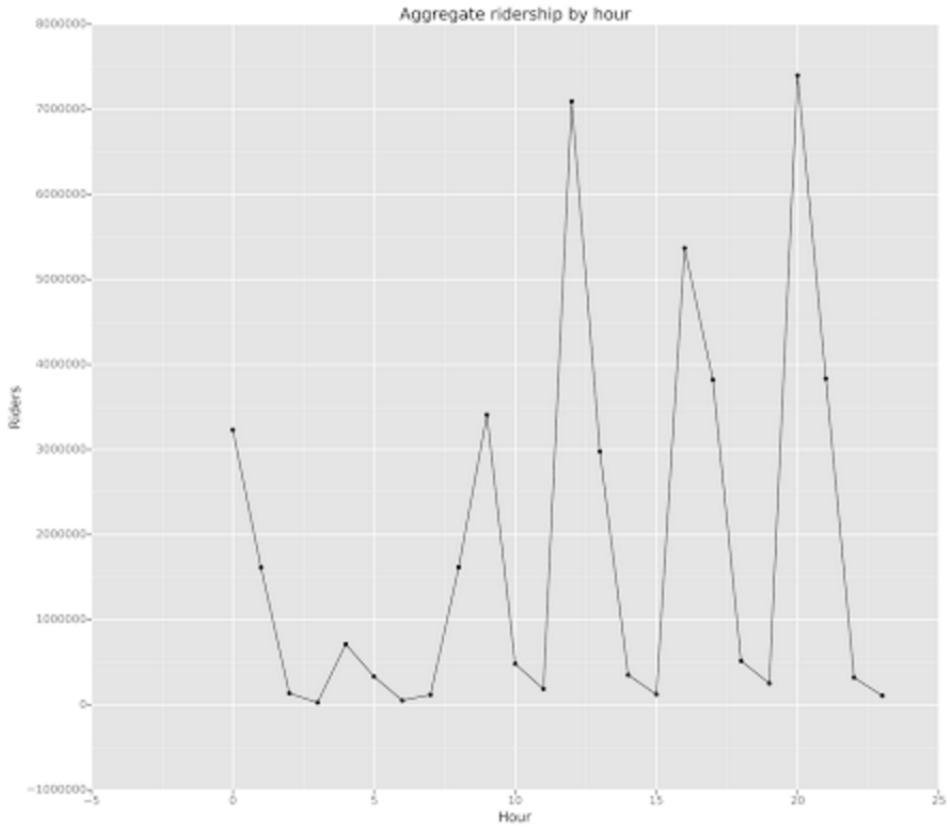
- R2 values will be between 0 and 1. 0 indicates bad; 1 is good. My R2 value of roughly 0.5 does not indicate a high-quality result. This is a poor model.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

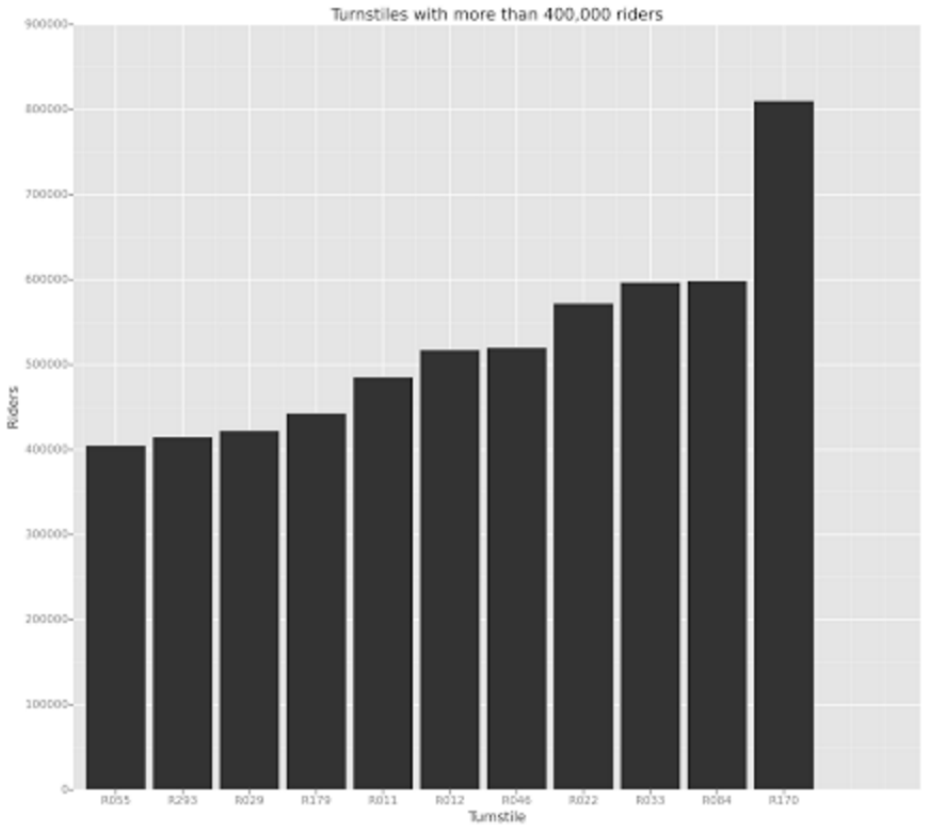
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

- Lesson 4: Data Visualization > Visualization 1



  - Ridership is high at rush hours, lunch, and again at 8 PM. (8 PM was a surprise to me.)

- Lesson 4: Data Visualization > 2 - Make Another Visualization



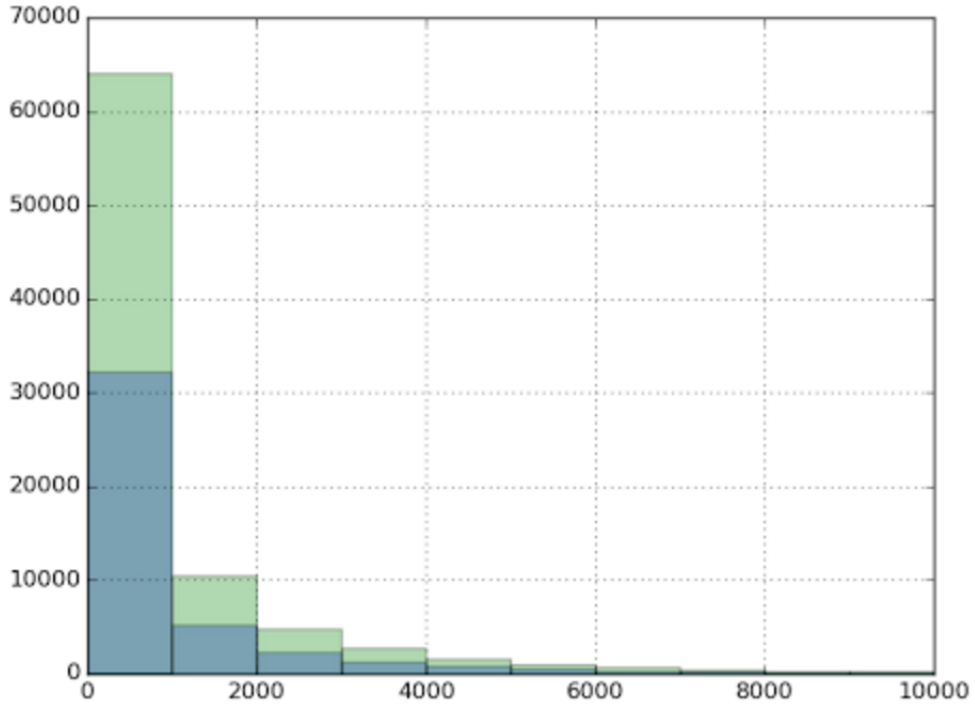
  - There are 11 turnstiles with over 400K riders in the period. Turnstile R170 is is much more popular than any other turnstile in the system.

3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

You can combine the two histograms in a single plot or you can use two separate plots.
If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.

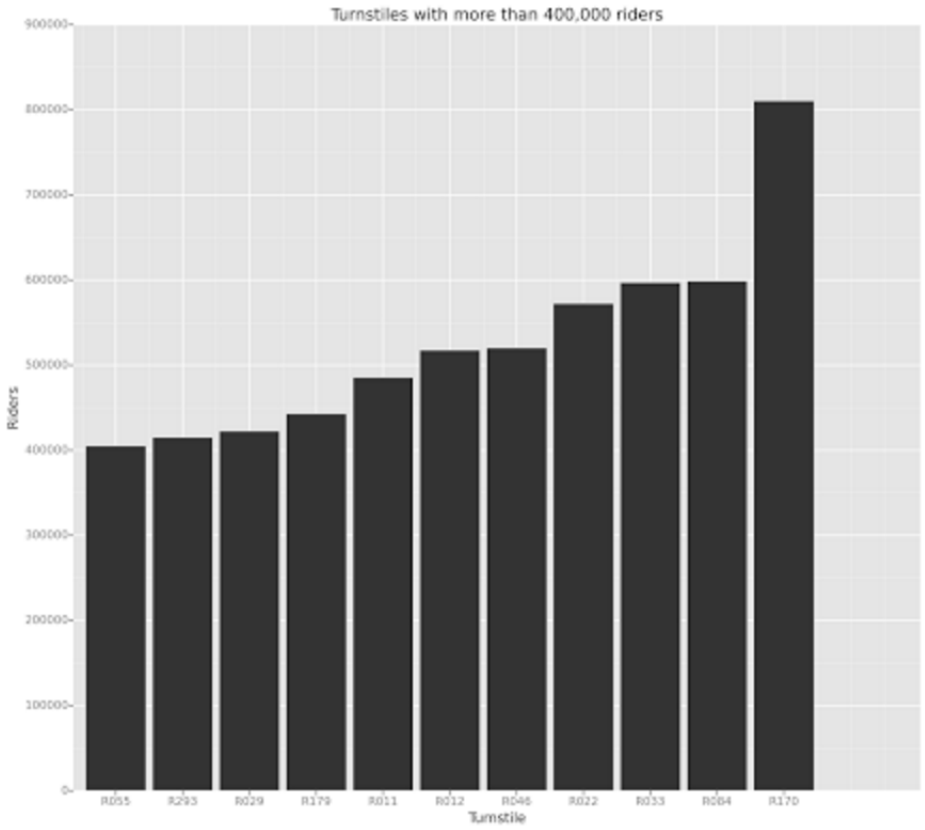
Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.

- Problem Set 3: Analyzing Subway Data > 1 - Exploratory Data Analysis



3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

Ridership by time-of-day
Ridership by day-of-week



Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

- Dry.

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

- The histgram in "Problem Set 3 Analyzing Subway Data > 1 - Exploratory Data Analysis" clearly shows more riders when dry.
- "Problem Set 3: Analyzing Subway Data > 2 - Welch's t-Test" shows that the means are statistically different, and the gradient descent implementation
- "Problem Set 3: Analyzing Subway Data > 5 - Linear Regression" calculated a linear regression (with a low quality).

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

Dataset,
Analysis, such as the linear regression model or statistical test.

- "Problem Set 3: Analyzing Subway Data > 5 - Linear Regression" calculated an R-squared value indicating poor quality.
- Calculating regression model via Gradient Descent is simple to implement, but the course material explained that there are more accurate methods.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?