Analyzing the NYC Subway Dataset

Questions

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, 4, and 5 in the Introduction to Data Science course.
This document addresses part 2 of the project. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

Section 0. References

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

Section 1: Statistical Test

I used this book as a general reference on statistical methodology:
Herbert Basler: Grundbegriffe der Wahrscheinlichkeitsrechnung und Statistischen Methodenlehre, Physica Verlag, 11. Auflage
Mainly Chapter 3, especially 3.4.2 (t-test) and 3.4.5 (Mann Whitney U-test)

On z-transformation:

http://de.wikipedia.org/wiki/Studentisierung

why chose a significance level of 0.05:

http://www.p-value.info/2013/01/whats-significance-of-005-significance_6.html

http://www.radford.edu/~jaspelme/611/Spring-2007/Cowles-n-Davis_Am-Psyc_orignis-of-05-level.pdf

Interpretation of p-value:

http://stats.stackexchange.com/questions/46856/interpretation-of-p-value-in-hypothesis-testing/46858#46858

Section 2: Linear Regression

Plotting residuals

http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

http://matplotlib.org/users/pyplot_tutorial.html

http://matplotlib.org/api/pyplot_api.html#matplotlib.pyplot.hist

computing R^2
http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html

http://docs.scipy.org/doc/numpy/reference/generated/numpy.sum.html

http://en.wikipedia.org/wiki/Gradient_descent

http://en.wikipedia.org/wiki/Coefficient_of_determination

Correlation matrix to find out the best candidates for linear regression:

http://pythonprogramming.net/pandas-statistics-correlation-tables-how-to/

http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.corr.html

Interpretation of R^2
http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
http://people.duke.edu/~rnau/rsquared.htm

Section 3: Visualization
Introduction and overview with good examples:
http://opr.princeton.edu/workshops/201501/ggplot2Jan2014DawnKoffman.pdf

overview on existing geoms:
http://docs.ggplot2.org/current/
http://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

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?

I used the Mann Whitney U-test to find out if there is a significant difference between the number of riders with rain versus the number of riders without rain.

The Mann Whitney U-test is used to test if, analyzing two input variables x and y with unknown distributions, one of them generates higher or lower values at random draws than the other one.

The null-hypothesis states that H_0 : P (x > y) = 0.5, i.e. the probability to draw a higher value is the same for both populations x and y.

Transferred to our question here: the null-hypothesis says that there is no increase in the probability to observe more subway riders, may it rain or not.

I used the two-tail p-value, because the question in test aims at differences in both directions, there could as well be more riders or less riders when it rains.

I choose a significance value of 0.05. There is no deterministic reason to pick exactly 0.05 but this has established itself as an agreed level of significance following the idea that an event which occurs 5% of the time or less is rare enough to assume that there is a different cause but mere chance behind it.

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.

Welch´s T-Test postulates that the data set follows a normal distribution. But the histogramm has clearly shown that this is not the case. Therefore Welch´s T-Test is not applicable on this set.

However, the histogramm has shown that the Entries at Rain and the Entries at no Rain follow a very similar distribution. This makes it a candidate fort he Mann Whitney U-test (see assumption above in 1.1., second paragraph).

It has to be considered that the number of without rain occurencies is much higher than the number of with rain occurencies. Therefore the test to choose has to be robust against different sample size. This is true for the Mann Whitney U-test. This is why I picked this test for this data set.

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.

The mean values between with rain and without rain differ by 1105.4 - 1090.3 = 15.1, which means that the average number of riders with rain is about 15.1 higher than the number of riders without rain.

The question is: can this difference be explained by chance or is it a significant difference? This question shall be answered by a statistical test, in this case the Mann-Whitney U-Test.

Executing this test produces the result of p = 0.024999912793489721, which is the value for the one sided test.

1.4 What is the significance and interpretation of these results?

The result of our test shows a p-value of 0.024999912793489721, which must be doubled for the two sided test and leads to a value of 0.04999982558697944. Compared with our alpha-Value of 0.05 we see that p < alpha.

Based on our siginificance level of 0.05 we have to reject the null-hypothesis that there is no difference between the number of riders with rain and without rain. Thinking of 'Fisher's disjunction': Either a rare event has happened or the null hypothesis is false. Following our approach we think that the event is rare enough to reject the null hypothesis.

The result is that significantly more people use the Subway when it rains.

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:

1. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

I ussed the the Gradient descent approach.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
I used the features 'Hour', 'meanwindspdi' and 'mintempi'. The dummy variable "UNIT" was also added.

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.

- Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.”
- Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R^2 value.”

I calculated the correlation coefficient for each variable in relation to the variable ENTRIESn_hourly:

(print(turnstile_weather.corr()))

The result:

	correlation coefficient to ENTRIESn_hourly
Hour	0.175430
ENTRIESn_hourly	1.000.000
EXITSn_hourly	0.744316
maxpressurei	-0.017084
maxdewpti	-0.009893
mindewpti	-0.020135
minpressurei	-0.020517
meandewpti	-0.016198
meanpressurei	-0.016128
fog	0.011368
rain	0.003062
meanwindspdi	0.026627
mintempi	-0.029034
meantempi	-0.022796
maxtempi	-0.014303
precipi	0.009665
thunder	NaN

Taking only the absoulute values into account it shows that the two highest values are ENTRIESn_hourly and EXITSn_hourly. This is trivailly true and self evident. It proves that the approach produces correct results but does not reveal any insight.

The next three variables are hour, meanwindspdi and mintempi. Before picking this values one should ask if they seem to make sense judging by intuition: Hour makes sense, otherwise there would be no such thing as a rush hour. Mintempi (negatively correlated) makes sense, because people should be likely to prefer the train instead of walking when temperature decreases (at least in New York, there might be a different situation in the desert). What about meanwindspdi? It is positively correlated to the number of entries. The more wind, the more people riding the train? When I think of women being afraid to ruin their hair-do: It makes sense for me.

So the three variables which showed the highest correlation to the entries (without being trivial) made sense according to my intuition. This is why I picked them for my model.

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
-1.57248270e+02 (hour), -1.55110529e+02 (meanwindspdi), 1.21866703e+03 (mintempi)

2.5 What is your model's R^2 (coefficients of determination) value?
I calculated an R^2 Value of 0.464610132035.

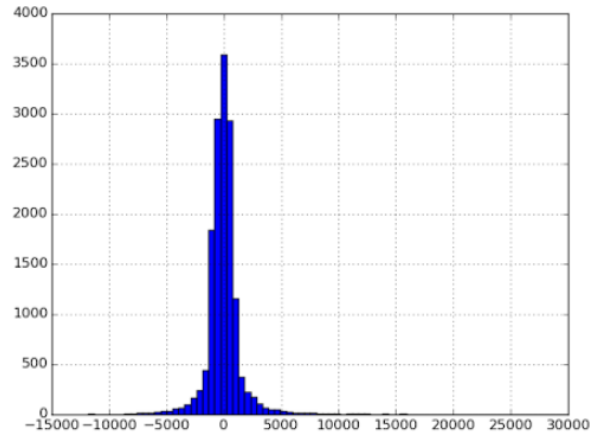
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

R^2 is a measure indicating how well a a statistical model fits to the actual data, where 0 is the worst and 1 is the best possible value. R^2 expresses the proportion of the original variability the model is able to explain. So our value of 0.46 tells us that 1- 0.46 = 54 % of the initial variability is left unexplained by our model. This looks like a pretty bad result.

Most authors state that a high R^2 does not necessarily mean that the model is good. But a low R^2 value is never an indicator for a good fitting model, even if in some cases the model might still be helpful to get some insight into the data.

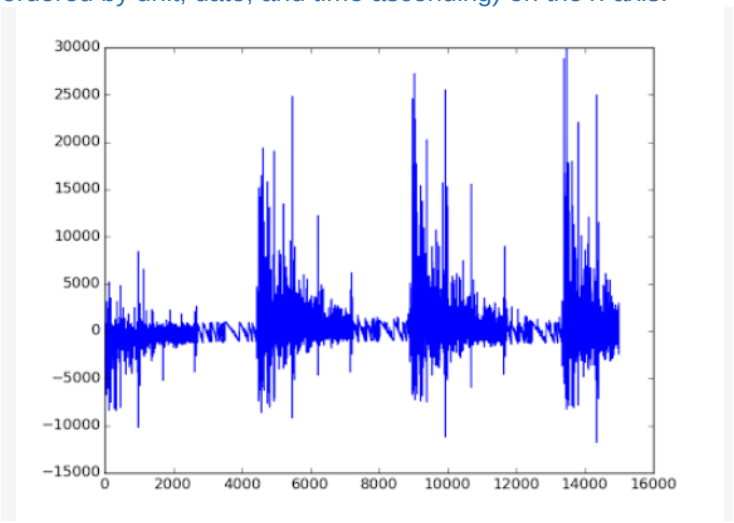
This is the time to have a look at the residuals distribution. The residuals as the difference between calculated and actual value should be distributed randomly, meaning that they follow a normal distribution. The analysis of the residuals tells a lot about how well the model fits to the data.

First let us have a look at the histogram to get an idea of the distribution:



Even if this curve looks symmetric around the 0-Value and to some degree reminds of a bell-curve, it strikes that this curve shows long tails, so our linear model unfortunately seems to produce a number of very large residuals.

So let us have a deeper look at that residuals and see if there is any pattern we can identify. The following plot shows the residuals on y-axis and the values by order of occurrence (which means ordered by unit, date, and time ascending) on the x-axis.



We see very clearly that there is a strong cyclic pattern, showing three big and one smaller peak. With this we have identified a basic relationship in the data which is non-linear. This is a structure which a linear model will not be able to reflect.

We can conclude that a linear model is not the appropriate type to represent the number of riders while a non-linear model would be more promising.

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

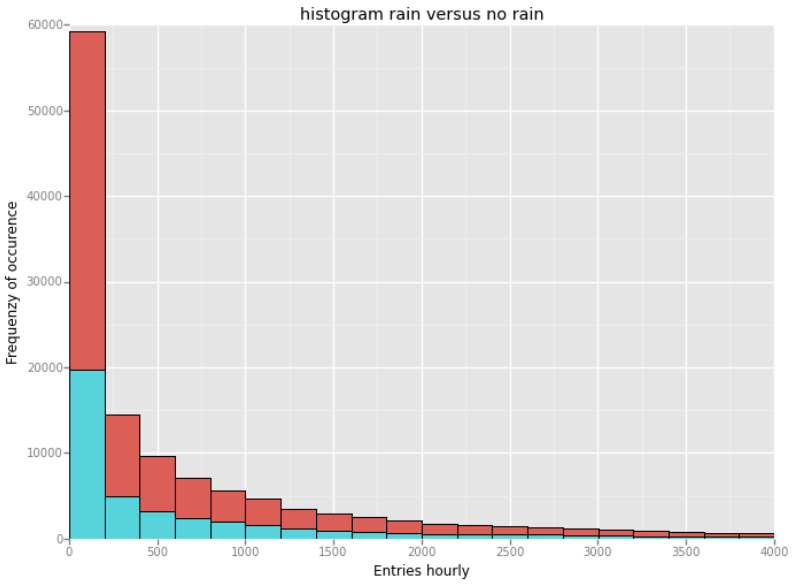
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, 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.

My solution:

```
plot = ggplot (aes(x='ENTRIESn_hourly', fill = 'rain'), data = df )
plot + stat_bin(binwidth = 200 , color = 'black') \
+ xlim(0, 4000) \
+ ggtitle ("histogram rain versus no rain") \
+ xlab('Entries hourly') \
+ ylab('Frequenzy of occurence')
```

.. creates this plot:



Blue: rain

Red: no rain

(I could not add an automatic legend in the plot, it seems this feature is not fully implemented in Python.)

This plot shows the hourly entries and their frequenzy of occurence. It can be seen that the frequenzy without rain exceeds the frequenzy with rain for every number of hourly entries, which is not very surprising.

We can further see that the distribution is extremely skewed. The frequenzy of low values with <= 200 entries per hour exceeds the accumulated rest of all entries. This tells us that far most of the overall traffic volume takes place as a „base load“ with a low number of passengers entering the train at a time.

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

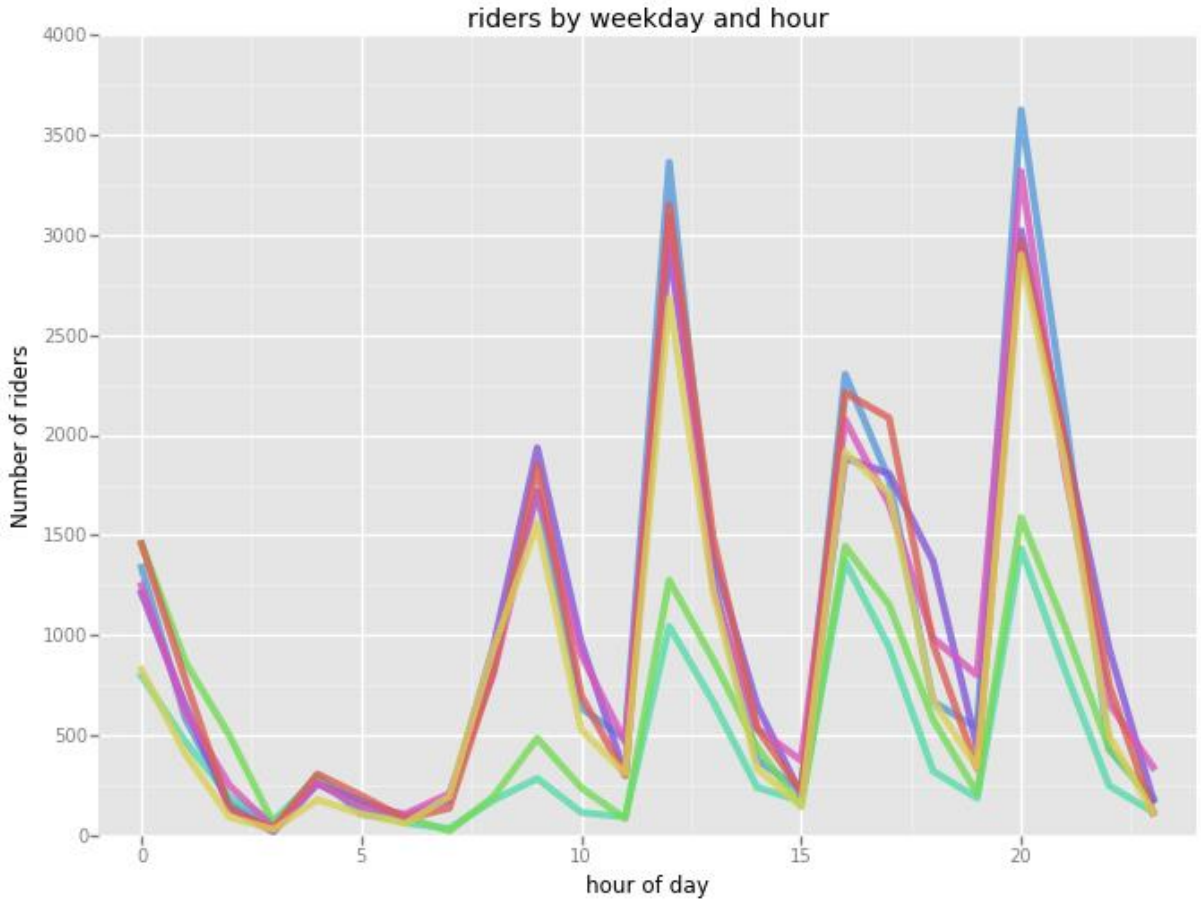
After adding the weekday to the data set:

```
from datetime import datetime
df['Weekday'] = map(lambda dat: datetime.strftime(datetime.strptime(dat, '%Y-%m-%d'),
'%A'), df['DATEn'])
```

.. I used the following code:

```
p = ggplot(aes(x='Hour', y='ENTRIESn_hourly', color = 'Weekday'), data=df)
p + stat_summary(geom='line', size = 4, alpha = 0.8) \
+ xlim(-1, 24) \
+ ggtitle ("riders by weekday and hour") \
+ xlab('hour of day') \
+ ylab('Number of riders')
```

.. to generate this plot:



The outliying green lines represent Saturday and Sunday, the other colors stand for the rest of the week.

This visualization allows us to compare the distribution by time of day for the different weekdays. The idea is not to look at the exact values but to get an idea of the patterns in the data.

With a rough visual examination we find that there seem to be two types of curves: weekend versus working days (this is a conclusion drawn out of the visualization, not an input, therefore in my opinion it makes sense to show the weekdays with separate lines, despite they look very similar) .

Again we might ask ourselfes: does this result make sense? Of course it does, human behaviour obviously differs a lot between weekend and working days.We find that the number of riders on Saturday and Sunday is lower than the number for any of the working days, which makes sense.

We see that there are four peaks each day around 9h, 13h, 17h and 20 h. However, these peaks are no rush-hours, as one might expect at a first glance. If we have a look at appendix 1 we see a pattern in the hourly reports in the data. Typically there are 3 hours skipped. The most common patterns are 1, 5, 9, 13, 17, 21 and 0, 4, 8, 12, 16, 20. This is what we see here as well.

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?
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 linear regression I conducted did not provide a valid result. The R^2 value of 0.46 showed that the model was only able to explain 46 % of the variability of the entries variable, which is a low value. So this linear model was not able to explain the number of riders based on linear relationships based on the chosen features.

The analysis of the residuals showed the reason for this. There are strong cyclic processes at work which make a linear analysis inappropriate to reflect what is going on.

The statistical test was much more use in this case. Having detected a higher average number of riders with rain compared to no rain, I used the Mann-Whitney U-Test to find out if this difference is statistically significant. The result was that significantly more people use the train when it rains.

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:

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

1. Observations on the Dataset

Since there is no such thing as a definition of the data set given, it is left to the flair of the analyst to figure out how the dataset is composed.

There are a couple of observations to be made:

- The time variable TIMEn usually follows a pattern where three values are skipped, for example 1, 5, 9, 13, 17, 21 or 0, 4, 8, 12, 16, 20 (see appendix 1).
- Knowing that each line in the data set represents more than one hour: Could it be that “ENTRIESn_hourly” is an average “hourly” value? No, because the number is given with one digit after the decimal point and we can see that all values are integer. This would not be the case if they were average values.
- So there must be accumulated numbers in the “ENTRIESn_hourly” field. Next question: Are they accumulated since the last entry or accumulated for the whole day? In other words: is the counter set to zero after each entry or does it go on counting the whole day? The numbers show that “ENTRIESn_hourly” does not continually raise during the course of the day.
- So the result of our investigation so far is: the number of entries (ENTRIESn_hourly) represents all entries which have been accumulated up to that hour of day since the last hour-entry.
- Not all time series follow the described pattern. There are also other values (odd values like 14:11:12 or 7:20:01) in the data, which seem to occur randomly.
- There is a difference in the units here. Units with an ID of R540 or higher do not show the regular time series at all but only the odd values (see appendix 2). Sometimes theses records contain no entries, for example unit R540 on 1.5.2011 shows about 20 values in series with odd time values and an entry value of 0.

- The data represents the month of May 2011. The expected number of lines per unit should therefore be at least 6 * 31 = 186, because there are six hours in the regular pattern and 31 days of May. The last column in appendix 1 shows that the actual value falls below this mark a couple of times in the sample.
- Units with ID from till R464 are represented continuously, but there are gaps after that, the biggest one from 469 to 535. This could mean that units are missing completely.

So we get a feeling for the data and some possible problems:

- Not all units follow the same pattern of hourly values and we don't know the reason.
- Sometimes values are missing.
- Chances are that whole units are missing as well.
- There are disruptions in the number of entries per unit.
- There are records with 0 entries which make no sense and might indicate errors.

This shows us that there is a lot we do not know about our data and the circumstances under which they came into being.

Other general shortcomings of the dataset are:

There is only one month represented in the data. This is a sample which could lead to misleading results because we are only analyzing the effects of weather in May. For example the behaviour might be different in winter.

There might have been some special conditions in that paricular month which could have influenced the behaviour. For example, a strike of the taxi drivers would have deformed the data completely.

For me the most serious shortcoming is that the granularity might be too low to find out what we are looking for. As we can see in appendix 3, each day gets classified as rain = 1 or rain = 0 for the whole area of New York and the whole day.

This is a rough representation of the nature, in reality there will be a lot of graduations in intensity spanning from no rain at all over shower to cloudburst over time and place.

Capturing data only every three hours (four values per day) means another loss of information, but since the rain-information (one value per day) is even rougher this has no significant effect any more.

An ideal data basis would tell us if it was really raining at the place and the time when the passenger entered the subway.

2. Reflection of Analysis

The visualization has shown us that there is a difference in the behaviour of the subway riders between the working days and the weekend days. This means that we have cyclicity here.

There are two more cyclicities, one being the daily structure of traffic ("rush hours") and the other being a consequence of the special way the data is reported. We typically have a lag of three hours till the next record is generated.

This multiple cyclicities impeded the linear approach, which can not be a surprise: The very basic assumption for the use of a linear model is that there actually is a linear relationship in the data which can be detected. Important to notice: this is not a question of bad adjustment of the model. The problem is that the model is just not capable to reflect the behaviour of the riders.

The observations on the dataset gave us more than enough good reasons to be very cautious about generalizing the result of the statistical test despite the significance we achieved.

We should keep in mind that the reliability of the results depends completely on the quality of the input.

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

Appendix 1: number of records per unit and hour as typical for units R001 to R536 (sample).

UNIT	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Sum
R001		29				29		3		29			1	30	5	1		30				29			186
R002		30				30				30				30	3			30				30			183
R003	30				30			2	14	1	2		31		4		29				29				172
R004	30				30			1	20		1	2	32	1			29				30				176
R005	30				30			1	17		4		31		2		30				30				175
R006	29				29			6	30		4	2	30	1	1		28		1		29				190
R007	30				30				15	4	1		30		1		30				29				170
R008	30				30				13			1	30				30				30				164
R009	30				30				22	1			29		1		29				30				172
R010			29				30	2			30				30	1			30				30		182
R011	30				30		1	8	29		5	3	30		1		30				30				197
R012	30		1		30		1	3	32	2	1	6	35	4	4		30				30				209
R013	30	1			30		1		31	3		2	30	1		3	30				30				192
R014		30				30		3	1	30	1	1		30	2			30			1	31			190
R015			30			5	36	3		1	38			1	48				37				30		229
R016	30				30			1	28		3	1	30	1	3		31				30				188
R017	30				30		1	1	30			1	33	3			30				30				189
R018	30				30		1	2	36	5			32	7	8	2	31		2	4	30		4	7	231
R019	30				30				30	3	3	5	32		5	1	31				30				200
R020	30				30		2	7	32	2	2		31	3	3	3	31	1			30				207

Appendix 2: number of records per unit and hour for units R540 to R552 (complete)

UNIT	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
R540	189	191	201	188	188	198	196	186	187	181	171	178	174	166	182	175	183	180	178	166	189	188	192	193
R541	256	250	252	258	255	251	262	243	248	242	222	226	235	245	250	244	253	250	238	249	238	244	259	252
R542	58	67	66	63	64	61	63	66	61	66	61	51	60	62	61	61	63	59	56	62	62	55	66	65
R543	173	174	179	183	172	177	182	186	159	162	161	159	163	172	180	175	168	177	172	179	170	169	173	181
R544	80	89	92	79	86	89	86	90	77	79	90	76	75	92	84	80	79	88	80	88	86	75	91	90
R545	77	73	85	75	72	86	73	77	81	73	72	81	66	73	80	71	74	85	69	72	81	70	73	82
R546	72	82	76	79	77	73	79	80	73	73	64	62	65	73	66	77	80	63	72	81	65	72	81	71
R547	37	35	30	39	35	30	39	35	31	39	30	29	30	32	30	39	32	31	37	32	33	38	34	30
R548	29	28	26	31	25	29	28	28	31	30	24	22	17	18	22	35	29	22	27	23	23	27	23	26

R549	498	552	527	474	542	548	488	523	511	485	490	446	437	530	522	462	529	531	506	521	511	496	542	527
R550	293	297	298	280	284	313	293	270	285	273	265	288	280	284	286	286	283	295	294	281	279	296	295	283
R551	134	127	137	133	126	135	135	125	126	130	113	119	138	121	137	139	122	128	137	131	133	135	125	134
R552	145	152	159	144	157	170	147	136	125	102	120	139	140	149	158	144	159	143	147	167	147	137	162	160

Appendix 3: granularity of rain-information

DATEn	rain	HourMin	HourMax
2011-05-01	0.0	0	23
2011-05-02	0.0	0	23
2011-05-03	0.0	0	23
2011-05-04	1.0	0	23
2011-05-05	0.0	0	23
2011-05-06	0.0	0	23
2011-05-07	0.0	0	23
2011-05-08	0.0	0	23
2011-05-09	0.0	0	23
2011-05-10	0.0	0	23
2011-05-11	0.0	0	23
2011-05-12	0.0	0	23
2011-05-13	0.0	0	23
2011-05-14	0.0	0	23
2011-05-15	1.0	0	23
2011-05-16	1.0	0	23
2011-05-17	1.0	0	23
2011-05-18	1.0	0	23
2011-05-19	1.0	0	23
2011-05-20	1.0	0	23
2011-05-21	1.0	0	23
2011-05-22	0.0	0	23
2011-05-23	1.0	0	23
2011-05-24	0.0	0	23
2011-05-25	0.0	0	23
2011-05-26	0.0	0	23
2011-05-27	0.0	0	23
2011-05-28	0.0	0	23
2011-05-29	0.0	0	23
2011-05-30	1.0	0	23

```
(
Source:
SELECT
Turnstile_data_master_with_weather.DATEn,
Turnstile_data_master_with_weather.rain,
Min(CInt([Hour])) AS HourMin,
Max(CInt([Hour])) AS HourMax
FROM Turnstile_data_master_with_weather
GROUP BY Turnstile_data_master_with_weather.DATEn, Turnstile_data_master_with_weather.rain
ORDER BY Turnstile_data_master_with_weather.DATEn, Turnstile_data_master_with_weather.rain;
)
```