Abstract:

Our analysis of the city bike data was centered around studying the difference between the average number of bike trips on weekdays vs. weekends.

Our hypothesis was: The average number of bike-trips for a weekday is higher than the average number of bike trips on a Saturday or Sunday. Our reasoning for the hypothesis was that there would be more bike rides when people are using the bicycles for commuting to and from work. The Null hypothesis was that the average number of bike-trips of a weekday in February is equal or less than the average numbers of bike trips in the weekend.

For our analysis we used citibike data for the month of February 2014 so I calculations and conclusions are based on that dataset.

We defined
WEEKDAY = MONDAY through FRIDAY
WEEK-END = SATURDAY AND SUNDAY
BIKE TRIP = A pick up of a bike
AVERAGE NUMBER OF BIKE-TRIPS = Total number of bike-trips of "n" days divided by "n"

Significance level a=0.05

Our results are that on average there were 7,776.7 bike trips during a weekday but 8,650 trips on average on a weekend day. However the standard deviation of each was quite large (3,544 and 5,042 respectively).

We ran a two sample t-test on our samples and with a p-value of 0.60 we were unable to reject the null hypothesis of fewer trips on weekdays than weekends.

Data:

We cleaned and set the data so that we could visualize just the date time.

-
  ```
  data.drop(['tripduration', 'starttime', 'stoptime', 'start station id',  'start station name', 'start station latitude', 'start station longitude', 'end station id', 'end station name','end station latitude', 'end station longitude', 'bikeid', 'usertype',  'birth year', 'gender'], axis=1, inplace=True)
  ```

Then we grouped by days of the week of the data and we ploted in order to have the an idea of what the data looks like.

-
  ```
  data ['weekday'] = data ['date'].apply(lambda x: x.weekday())
  ax = ((data['date'].groupby([data['date'].dt.weekday]).count())).plot(kind="bar", color='IndianRed', alpha=0.5)
  ```

-
  ```
  tmp = ax.xaxis.set_ticklabels(['Mon','Tue','Wed','Thu','Fri','Sat','Sun'], fontsize=20)
  ```

-
  ```
  weekendday = []
  ```

- weekendday = data[(data['weekday'] == 5) | (data['weekday'] == 6)]
- weekday = []
- weekday = data[(data['weekday'] == 0) | (data['weekday'] == 2) |
  (data['weekday'] == 3) | (data['weekday'] == 4)
  | (data['weekday'] == 1)]

Then we count how many trip per day and we saved plot the total number of trip per day on the day.

- datelist = pd.date_range(start= '2014-02-01', end= '2014-02-28').tolist()
- countweekday = []
- for k in datelist:
    - countweekday.append(len(weekday[(weekday['date']> k) & (weekday['date'] < k+1)]))
- countweekendday = []
- for k in datelist:
    - countweekendday.append(len(weekendday[(weekendday['date']> k) & (weekendday['date'] < k+1)]]))

- pl.scatter(range(0,len(countalldayframe)), countalldayframe['occurance'],)
- pl.xlabel('Days of Feb')
- pl.ylabel('Trip per day')

Finally we calculated the mean and the standard deviation of the two samples, the weekday trip sample and the weekend day trip sample, and ran a two sample t-test on our data.

- stats.ttest_ind(countweekdayframe, countweekenddayframe, axis=0, equal_var=True)
- **Result:**
    - Ttest_indResult(statistic=array([-0.52148881]), pvalue=array([0.60644043]))