A clear statement of the goal of the mini-project (in short sentence(s) and most preferably in the form of (a) question(s) you intend to answer).  
-A brief summary of the approach you have followed to achieve your goal.   
-A brief introduction to the problem (what has been done before in this context, why is the problem you are tackling relevant, etc...).  
-Comprehensive descriptive analysis of the data (which variables you have available, are they continuous, discrete?, etc...).  
-Comprehensive presentation of results (which statistical models/techniques have you used -if any- and why they were adequate to tackle your problem with the data you had).  
-Conclusions and discussion.  
-Possible future work.

To visualize the profitability differences between each area we have used interactive maps provided by the leaflet library. We used this dataset (<https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6>) provided by Chicago Data Portal to visualize the borders between community areas. The differences are highlighted by attributing a color to each area, which indicates the size of the variable relative to the other areas. Our goal was to get a clear understanding of how area and time of the day affects revenue in each area, and being able to communicate these findings as clear as possible.

As a basis for this analysis we have plotted a map centered around Chicago. This was done by using the Leaflet library to extract a map over Chicago through Google’s map API. Thereafter we extracted *pickup\_community\_area*, *community\_name* and *geometry* from the dataset of community borders. *pickup\_community\_area* is a discrete variable uniquely identifying each area and *community\_name* is a string variable which contains its corresponding name. *geometry* is a Polygon(?) data type which essentially is a list of coordinates representing the borders of each area.

The map and borders between the areas form the base layer for the following visualizations.

We will first look into how total revenue is distributed among the areas. The relevant variables we will use are *pickup\_community\_area* and *trip\_total*. *pickup\_community\_area* is identical to its corresponding variable in the area dataset. *trip\_total* is a continuous numeric variable (?) representing the entire cost of the taxi trip for the passenger. We summarized over all *trip\_total* for each area, and colored each area by the result.

Since there are great differences in revenue between the areas we made the bins on a logarithmic scale. Because revenue between areas often vary by several orders of magnitude, the logarithmic scale gives a meaningful overview without concealing important information.

From the map it is clear that a few areas generate a disproportionate amount of the revenue. Does this imply that these are the most desired areas for each driver to be in? Not necesarrily. For each area there are several factors which vary; population, movement, amounts of taxis, etc. This means that although an area generates the most money overall, it’s not necessarily the same area which will earn the individual taxi driver the most money.

To not get muddled by these hidden factors in total revenue, we can take a different approach. We will look at the average time from a taxi drops someone off in an area, until the same taxi picks someone else up in the same area. We will only consider pickups and drop-offs which happen in the same area. This is because taxis which travel from one area to another between drop-off and pickup will inflate the average time. This could potentially lead to the central areas getting higher averages, as taxi drivers are inclined to travel back to the city center after a drop-off.

We extract *taxi\_id*, *trip\_start\_timestamp*, *trip\_end\_timestamp*, *pickup\_community\_area* and *dropoff\_community\_area* from our dataset. *taxi\_id* is an unique string identifier for each taxi and *pickup\_community\_area* and *dropoff\_community\_area* are identical to the previous description of *pickup\_community\_area*. *trip\_start\_timestamp* and *trip\_end\_timestamp* are date strings one the format “MM/DD/YYYY hh:mm:ss”, which indicates time of trip start and end. We should note that the timestamps has been rounded to nearest 15 minute for privacy purposes. Therefore the result will not be 100% accurate, but it will give a good estimate of the true time to pickup.

When considering optimal areas for each individual taxi driver, the plots paint a fairly different picture. The revenue plot and activity plot both point towards the areas around Loop as generating the most money. This is considered the center of downtown. In addition O’Hare, the airport of Chicago, is one of the top earners. When plotting the waiting time, however, it is clear that these taxis in these areas also are more prone to waiting time – with O’Hare averaging 95 minutes in waiting time.

None of these plots paint the whole picture on its own. Although the waiting time at O’Hare airport is large, a trip from the airport to the city center would generate much more revenue than a trip from Loop to West Town. To take these differences into account, we will combine three factors which will make a measure for *expected hourly salary* in each area.

Our expected salary is made up of two main components: expected income and expected time spent: (expected income / expected time). Expected time spent will further be divided into expected waiting time and expected trip time. These estimators(maybe not call them this) are obtained by averaging over our dataset in similar fashion as done earlier. The distribution for each area is shown below: (if we bootstrap, mention it here)

--- Wrap up ---

If not: write abot how earnings would be a natural next step for further analysis. This could in the end culminate in a machine learning model recommending each taxi driver where to drive after a pickup based time of day, time of week and current area.