Philadelphia 2017 - DSML Group Project

Indego

Author: Friedemann Ant (7354198), Jeremy Meidinger (7364610),
Maximilian Hachtmann (7364952), Maik Goebel (7370611)
Supervisor: Univ.-Prof. Dr. Wolfgang Ketter

https://github.com/Jerematix/dsml2021.git

Department of Information Systems for Sustainable Society
Faculty of Management, Economics and Social Sciences
University of Cologne

July 21, 2021

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Die Strafbarkeit einer falschen eidesstattlichen Versicherung ist mir bekannt, namentlich die Strafandrohung gemäß § 156 StGB bis zu drei Jahren Freiheitsstrafe oder Geldstrafe bei vorsätzlicher Begehung der Tat bzw. gemäß § 161 Abs. 1 StGB bis zu einem Jahr Freiheitsstrafe oder Geldstrafe bei fahrlässiger Begehung.

Friedemann Ant (7354198), Jeremy Meidinger (7364610), Maximilian Hachtmann (7364952), Maik Goebel (7370611)
Köln, den 21.07.2021

Contents

List of Figures

1 Data Collection and Preparation

The bike provider Indego offers bicycles for hire at various stations in Philadelphia. Indego's main goal is the optimal availability of enough usable rental bicycles at the required stations in order to provide every customer with a bicycle without having an oversupply of bicycles at the same time. For this reason, it is important to collect data and carry out resilient planning on the basis of this.
The given data set for our analyses from the city of Philadelphia in 2017 consists of two parts. On the one hand, we have data on all trips made in 2017. On the other hand, we have data on the hourly weather development in Philadelphia in 2017.

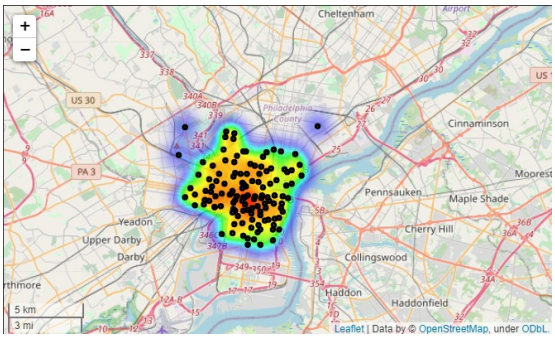
The **rides data** includes the start and ending time, as well as the used stations with their IDs. It also contains the riders user type. The data set contains six types of users. Four of these are subscription models for a year ("Indego365"), a month ("Indego30"), two days ("TwoDayPass") or a day ("OneDayPass"). In addition to these options, there is also the "WalkUp" user type, which is an ad hoc solution. The last subscription pass is the "IndegoFlex", which combines fixed and variable costs.
When working on the data we noticed, that there are a couple of outliers, that started before or after 2017 and some had unbelievable times. To gather valid data, we need to observe, whether this data might be just errors in data collection. So we tried conducting the 1.5 IQR (interquartile range) method, but altered it, because the upper limit was 44 minutes, which is a believable time. As a result we decided to take the 0.1 percent percentile and add the IQR to it, which comes out to around two hours. We decided to take this two hour border because nearly nobody will need two hours in Philadelphia to go from station A to station B. Because of this we thought that in most of the cases this should be a data error, which we take as outlier. We also eliminated some rides in which the ride duration was negative. After dropping this mistakes and outliers we added a column in which we record the duration of every ride, so that we just have to calculate it by once and not every time we need it.

We got a data set of hourly **weather data** in Philadelphia beginning in 2015 and ending in 2020. The data set gives us information about the maximum and minimum temperature in every hour as well as if there was precipitation. First we dropped all weather data, that is before or after 2017.
When looking at the data set, we noticed that some hours were occupied twice. Due to the ratio of these duplicates compared to the total number of data, as

well as the largely very small differences in the temperature information, we have decided to remove the duplicates. We also noticed that some hours were not recorded in the data set. In order to fill these gaps, we decided to average the temperature of the previous and the next recorded hour to get data for this missing hour. The previous hour is used as a basis for precipitation. We chose this type of data completion for the reason that there were no long times without any data acquisition, so we only had comparatively small gaps that we could validly fill using this average method. As a result of this we have ensured a comprehensive data frame. After the weather data had been processed correctly, we linked it to the ride data.

In addition to the rides and weather data, we also created **geo data** for the individual stations via Google Maps Geocode API Wrapper [1]. This gives us the possibility to create heatmaps. (see Figur No. 1)

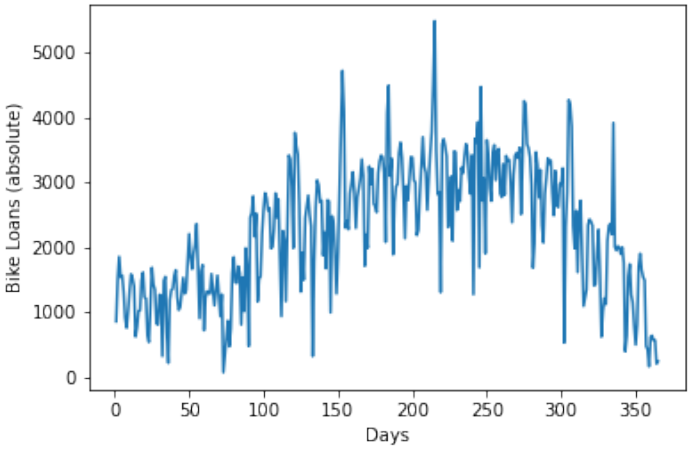
Figure 1: Heatmap geo data



2 Descriptive Analytics

The next step was to deal with the content of the corrected data.
In the first step, we look at whether we can determine demand patterns. In terms of rental behavior, we can determine that on working days there are on average 500 more loans per day than on weekends. Furthermore we can say, that in the hours from 8 am to 8 pm there are more bike rentals, with a peak at 4 pm. Both of these aspects mostly like belong to the fact that many employees use a rental bike. This also might be a aspect for the fact, that the stations in the middle of the city are used more frequently than the once at the outer city. When comparing the summer to the winter months we can see, that the number of bike rents in the summer month is much higher than the one in the winter month (see Figure No. 2), which goes along with our observation, that the bike usage falls the worse the weather is.

Figure 2: bike rents per year



After analyzing the demand patterns, we defined five KPIs in the next step to take a closer look at the detailed bike data.

We start with the analysis of the „**utilization**“ of the bicycle fleet. Here we look at the maximum hourly utilization of our entire bike fleet. To get the number of total bikes we added up the unique bike IDs and calculated a number of 1250 bikes. We then looked at how many bikes were rented per minute. We took the maximum of these 60 values per hour as the total value for the hour. In order to keep the presentation of the analysis clear, we have developed a code that displays the utilization of the requested month. As an example, we now compare the months of December (see Figure No. 3) and July (see Figure No. 4), which represent a very low and a higher utilized month respectively. In the comparison, we see that we could equip our fleet with fewer bikes in the winter months, which is relevant for aspects such as extensive maintenance, which will be taken into account in the following KPIs. In the summer months, on the other hand, we should make all bicycles available to our customers if possible.

Figure 3: July

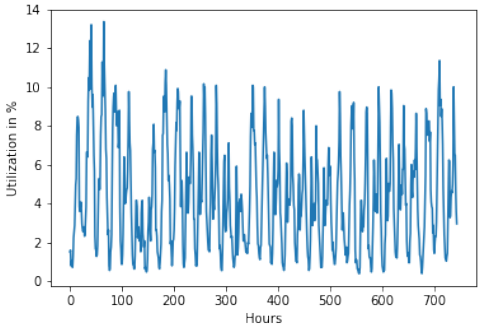
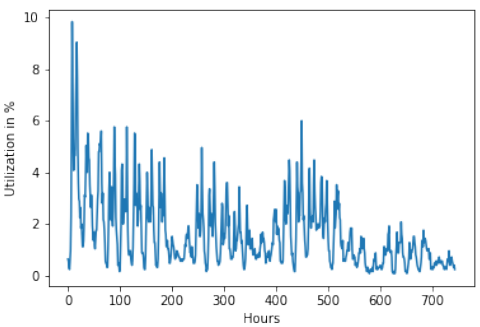


Figure 4: December



Next we look at the "**revenue**" we get by exceeding the gratis time. Unfortunately, we have no way of calculating the total income, because we have no information about how many users have the respective pass. So we can only check how many people exceeded their free minutes included in the pass. The Indego30 and Indego365 Pass bring 60 free minutes per trip. After this 60 minutes they have to pay 0.15 dollars for every minute more. Calculating these overdrafts resulted in income of 243,098.85 dollars for the whole of 2017 earned by the Indego30 Pass. The Indego365 Pass earns only 1,584.15 dollars. It might be worth offering the Indego365 free minutes limitless, thereby losing around 1,500 dollars, but generating more permanent revenue through the subscription. Unfortunately, we do not have any exact data for the One- or TwoDayPass on the payment models, but we have oriented ourselves on the "GuestPass" that still is valid in 2021, as this mirrors the OneDayPass. The GuestPass offers only 30 free minutes. These Passes generate 12,916.50 dollars (One) and 1,676.55 dollars (Two) which is a small amount on the whole revenue. The majority of income from overdrafts is generated by WalkUp users. This user has not subscribed to a passport and, according to a brochure from 2015 [2], which is unfortunately the only source of information about this war, pays 4.00 dollars per half hour from the first minute on. In total, WalkUp users generated revenues of 806,432.00 dollars in 2017. The last pass we have is the IndegoFlex Pass, which combines subscription and WalkUp. Here, owners only pay 2.00 dollars per half hour, but pay a fixed subscription amount. IndegoFlex users brought in a total of 28,592.00 dollars in 2017, which is a big difference to WalkUp users, but should still not be neglected completely. (see Figures No. 5, 6, 7, 8)

Figure 5: Q1

Figure 6: Q2

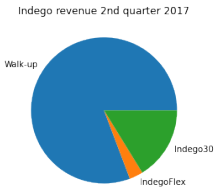
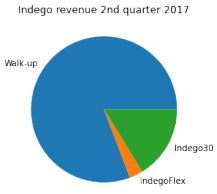


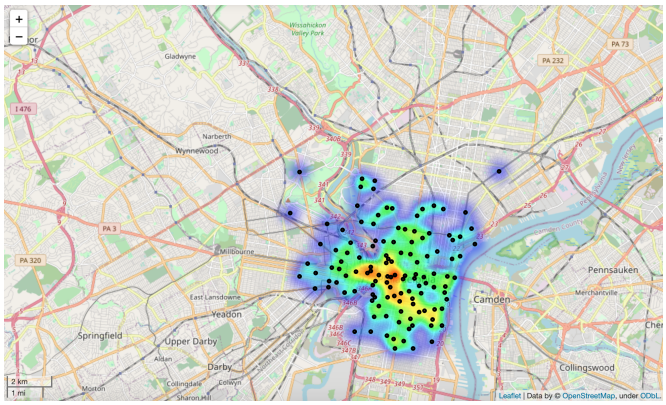
Figure 7: Q3

Figure 8: Q4



The “**bikes delta**” tells us how many bikes arrive at and driven off a station per hour. In the ideal case, we have a delta of 0 at every station, but this is not due to the fact that no trips are made, but that exactly the same number of bicycles arrive and leave. So we would not have to make any manual interventions, since all stations permanently supply themselves with "new" bicycles. Figure No. 9 shows the further you leave the city centre, the smaller the delta becomes. While the station "18th and JFK" has the largest delta with a delta of +5003 in 2017 and is located in the centre of the city, the deltas become smaller towards the outside, which means that bikes have to be transferred there manually by the company more often, e.g. from stations with a very high delta.

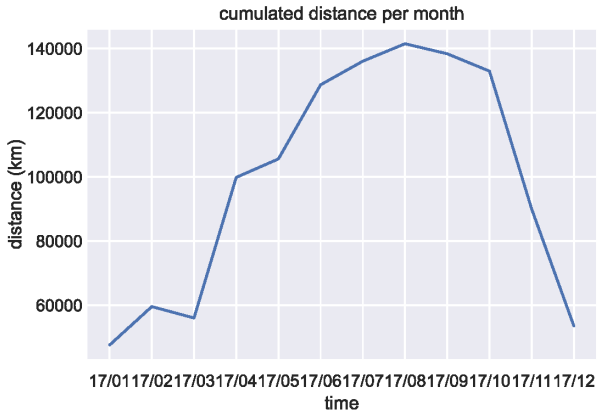
Figure 9: bikes delta



The fourth KPI “**distance**” tells us how much distance all bikes have covered. This information is especially important in order to effectively organize the maintenance. If a bicycle rides more distances, wear parts are worn out more quickly and thus maintenance is necessary. However, since it would be too extensive to look at each bike individually, we have decided here to record the sum of all trip distances per hour via the individual totals. In Figure No. 10 it can be seen that, as expected, the distances increase strongly in the summer months and rise to just over 140,000 km per month, while only a few km are covered in the winter months. However, the utilization rate considered in the first KPI recommends that we maintain the bicycles in the winter months, as they are not used as extensively there.

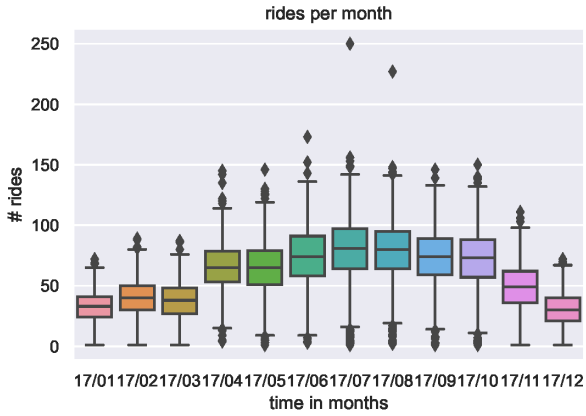
The fifth and final KPI deals with the "**rides per bike**". These have a similar background to the KPI “distance per bike” but deals with the number of rentals per month and not the total trip length. To visualize this, we have chosen a box plot, as this gives us a good indication of the average number of rides, but also of outliers. Here again we see that more rides take place in summer, but

Figure 10: distance



above all that there are much higher maximums than in the colder months.This enables us to differentiate between bicycles that are used a lot and those that are seldom used and thus gain important knowledge and information. It is also about the wear and tear of bicycle parts. However, the KPI also provides information on whether a bicycle is used less than average, which can be an indication that the bicycle is defective and should be controlled. (see Figure No. 11)

Figure 11: rides per bike



By combining all five KPIs, we can get a good overview of how much our offer is being used currently. For example, if there are currently a comparatively high or low number of trips, these trips are short trips to cover a short distance quickly, or long trips. Depending on the situation, we can react and, for example, adjust maintenance intervals variably. The KPIs also give us information about when we get the most profits from overdrafts over the total time. It must be ensured here that enough bicycles are available during these phases. On the basis of the calculated deltas to the individual stations, we can intervene here if necessary and counteract an out of stock by manual bicycle movements.

3 Predictive Analytics

In order to make predictions for the future and to be able to react to them, it is important to establish correlations between individual variables. For this purpose, we used various forms of regression analysis. Initially, we worked with values for whole days, but then switched to an hourly analysis due to the task at hand.

In the first step, we determined the first correlations by means of a **linear regression**. We decided to use linear regression because it is relatively easy to implement and is sufficient to quickly identify a rough correlation. We performed the first linear regression between the number of borrowers per hour (dependent) and whether the hour is on a day during the week (independent). The regression gives the result that if the hour is during the week, 21 more bicycles are borrowed per hour, which is significant at a very high level. However, it must be said here that the RMSE is very high, which means that the prediction can deviate greatly from the actual value by the RMSE. In addition, the R-squared value is very low and not near to 1 and therefore does not explain the variance well, which means that the model is not well fitted to the data.
The other linear regressions concentrate on the relation between the number of borrows per hours (dependent) and precipitation, temperature, moving weekdays and moving hours.
While all the linear regressions we ran are significant at a very high level, unfortunately they also all have a very high RMSE as well as a very low R-squared value, so the results must not be very meaningful. (see Figure No. 12)

Figure 12: linear regression

Linear Regressions for hourly data

Regression	Coefficient	RMSE	R^2
is_weekday	21,91	129,34	0,012
precipitation	-32,03	129,85	0,009
temperature	3,37	117,56	0,135
weekday	-3,36	127,86	0,006
hour	4,59	120,23	0,125

In addition to linear regression, we also looked at **multiple linear regression**. This has the advantage over linear regression that we can look at a dependency of more than one variable.
Here we used multiple linear regression to link the relationship between the num-

ber of borrowers per hour (dependent) and whether the day is during the week or not with the temperature (independent). The result of the regression was that a day during the week and higher temperatures mean more rentals per hour, but again the RMSE was relatively high and the R-squared value low, even if better than in the linear regression.

As a third and final regression, we used **polynomial regression** to represent well, for example, seasonal variations not well covered by linear using polynomial regression.
In this case, we first deal with the number of borrowers per hour (dependent) and the temperature in the corresponding hour (independent). The regression yields the result that higher temperatures mean more loans per hour. According to the R-squared value, it explains the variance only slightly, but still better than the linear regression.
We also ran a second polynomial regression to determine the relationship between loans per hour (dependent) and time of day (independent). It showed that most borrowing takes place between 8am and 8pm, with the number of borrowings increasing until 4pm and falling again from that point onwards. R-squared explains the variance slightly better than linear regression.

Figure 13: first degree

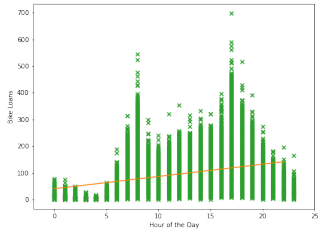


Figure 14: second degree

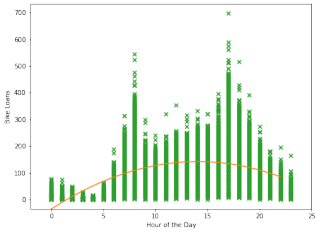


Figure 15: third degree

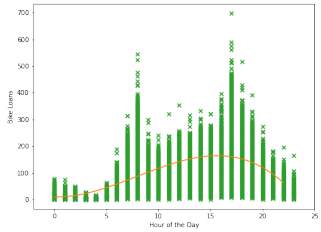
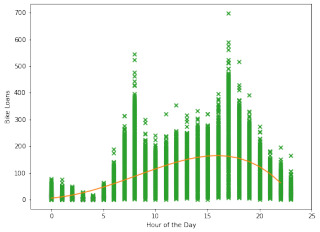


Figure 16: fourth degree



4 Conclusion

In summary, it can be said that an hourly forecast is very difficult, as very high fluctuations are possible. This can also be seen in the mostly high RMSE and low R-squared. If, on the other hand, the regressions are not done hourly but per day, the RMSE remains mostly high, but R-squared is much higher (about 0.6, which is unfortunately still far from 1), which makes the models basically usable. All in all, one can nevertheless conclude that the number of bicycle loans is high, especially in the summer months. Of course, good weather and no precipitation have a favourable effect here. In addition, the stations in the city centre are more utilised than those in the outer city areas. When it comes to monitoring the number of journeys and the kilometres travelled, one is also in a good position and can draw up a solid annual plan on the basis of this data.

5 Personal Contributions

5.1 Data Collection and Preparation

In the beginning we split up into several teams, which gathered and corrected the data. Jeremy and Maik worked on the rides data, while Friedemann and Maximilian worked on the weather data. Jeremy furthermore collected the geo data.

5.2 Descriptive Analytics

We collected ideas together in the group which KPIs we could use to work on. After choosing them we again split up to work on them:

1. Utilization - Maik
2. Revenue - Maik
3. Bike Delta - Maximilian
4. Distance - Friedemann
5. Rides per bike - Friedemann

5.3 Predictive Analytics

Jeremy and Maik together worked on the regressions.

Maximilian wrote the report based on the notes of the other group members.

References

[1] https://github.com/googlemaps/google-maps-services-python (accessed: 20.07.2021)
[2] https://nacto.org/wp-content/uploads/2016/04/2015$_{I}ndego_{I}ndego-Membership-and-Pricing_{B}rochure.pdf$(*accessed* : 20.07.2021)