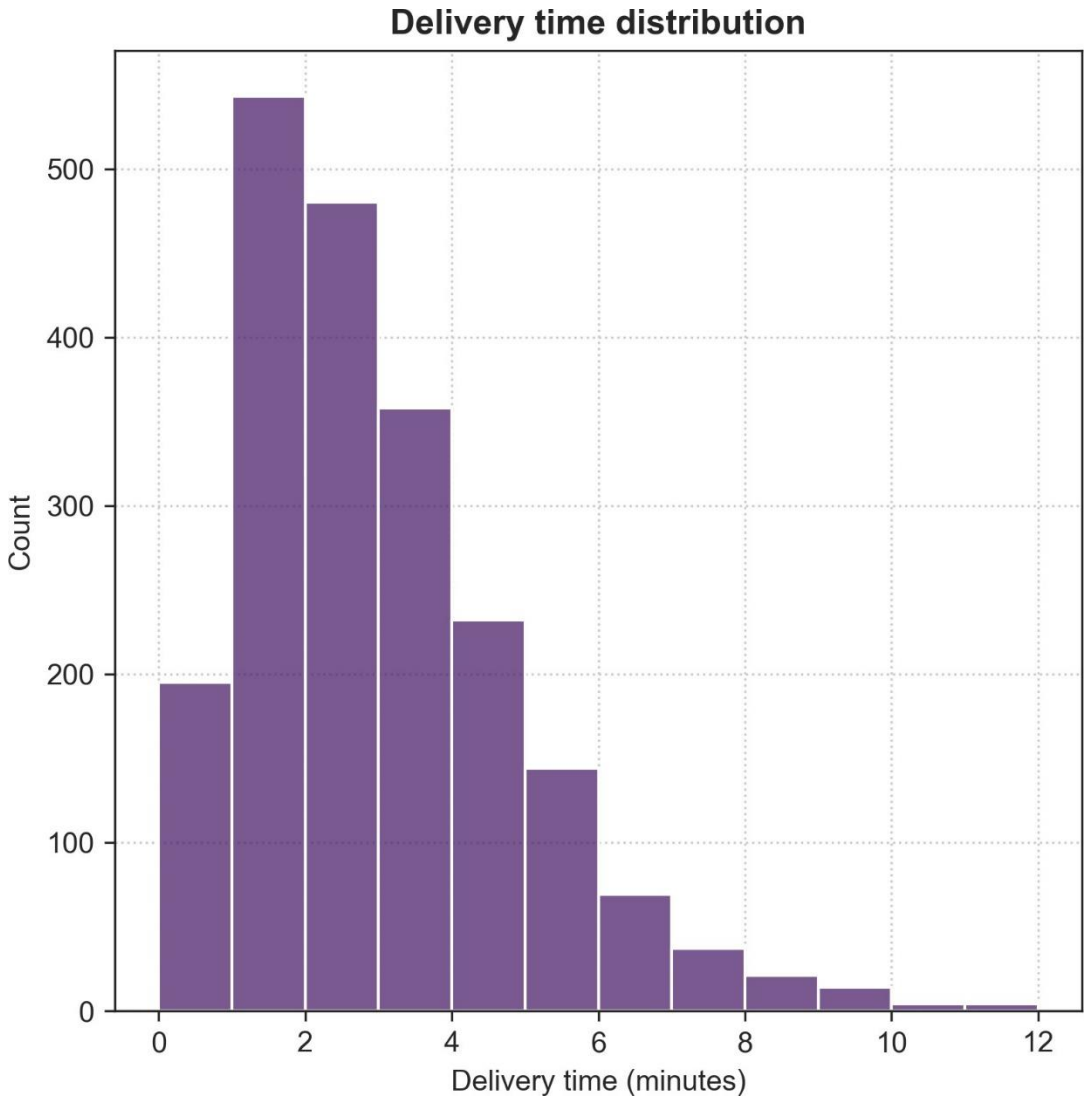
Jakub_Chłapek_analysis

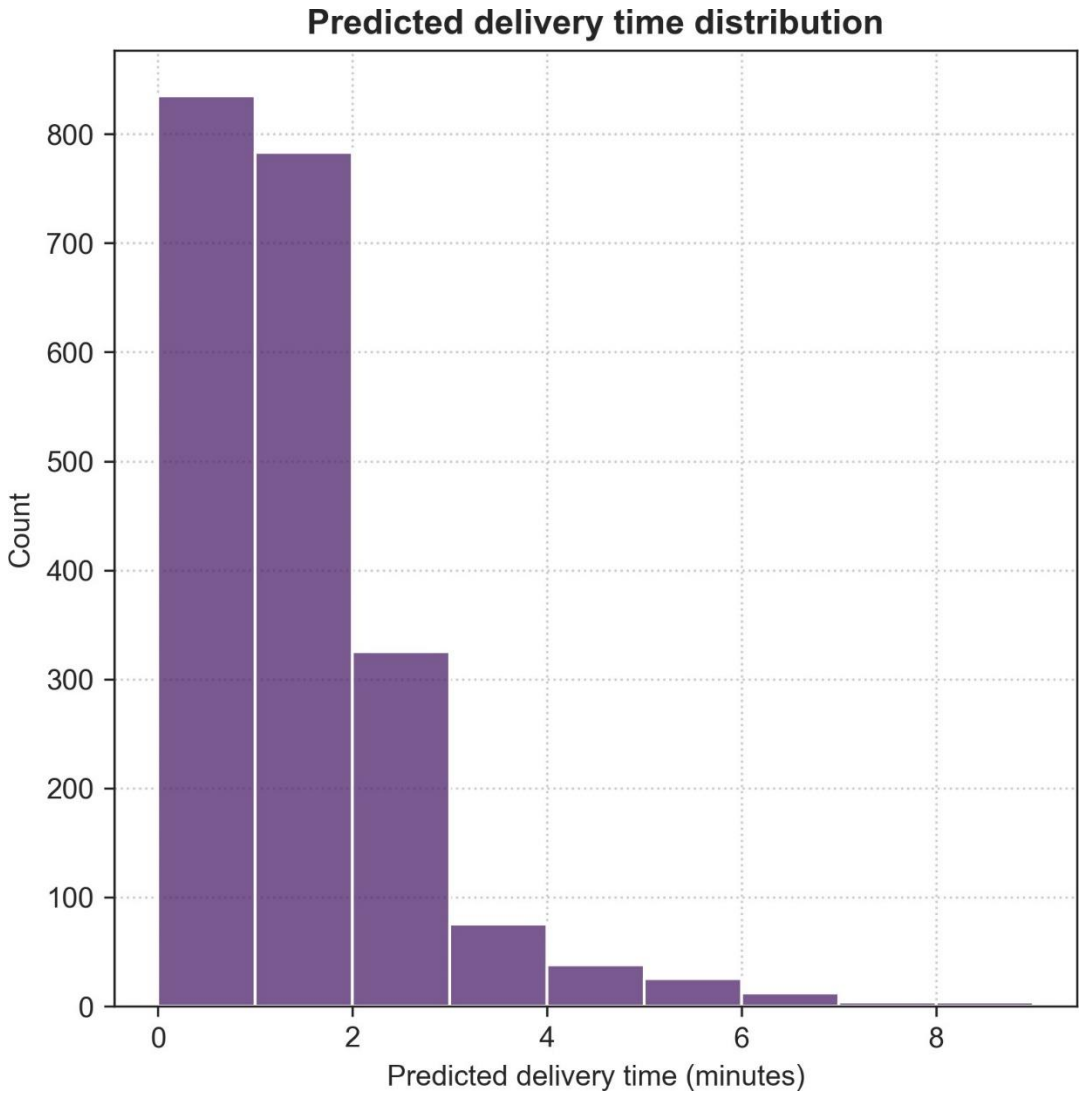
April 6, 2024

After connecting to the database I created earlier, I prepared tables for further analysis. I drop any duplicate rows from the data, and check for missing rows. In this case there are none that would be unexpected, so I move on.

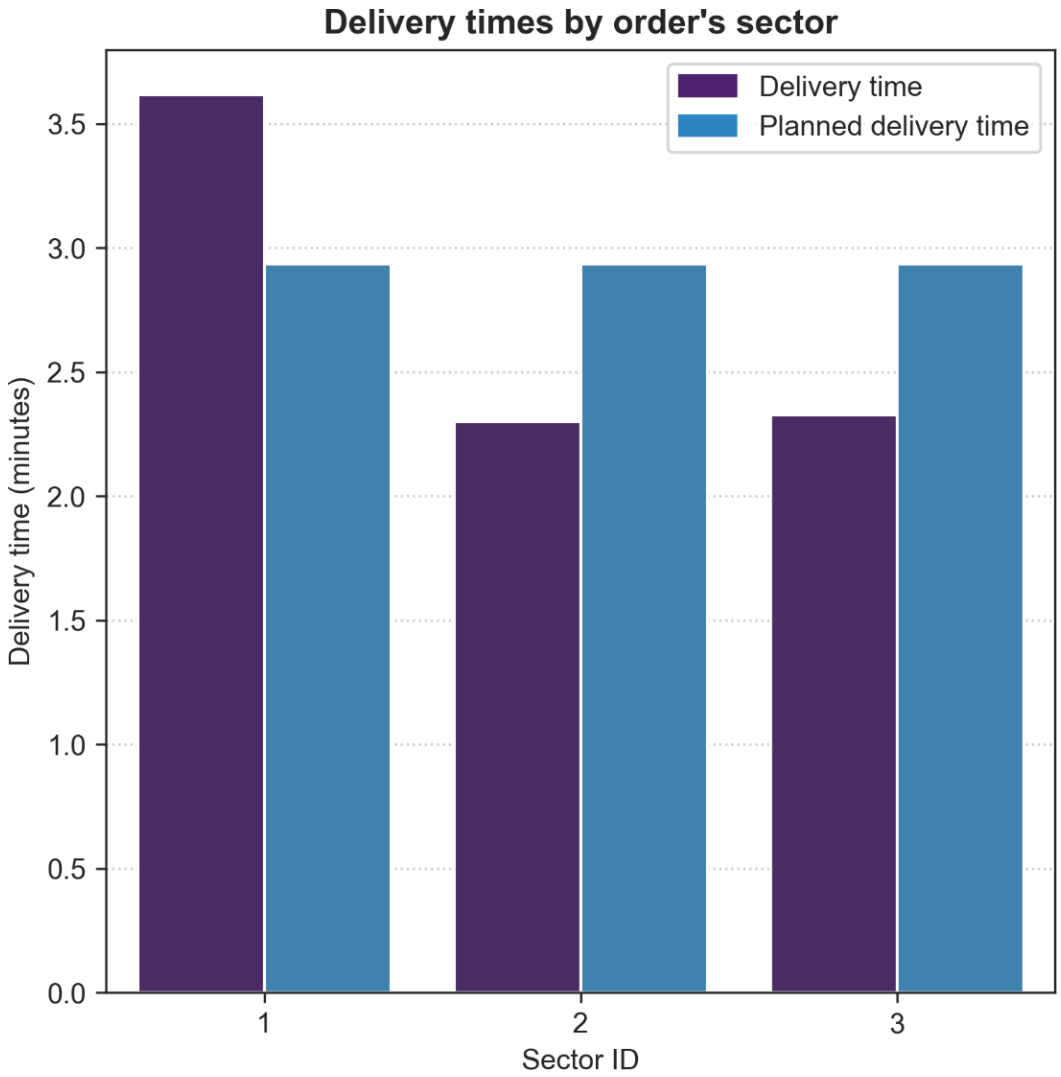
To create a delivery time histogram I use the route segments table. I removed outliers and some of the samples that had segment end times that were earlier than the start times. I also removed the ones with delivery time equal to 0, as I'm assuming that the delivery needs to take some time, and there was a considerable gap between the 0 delivery time records and the next lowest ones.



To create the prediction error, I merged the orders and route segments tables on their common ID key. I defined prediction error as the absolute difference between the planned delivery duration and the actual delivery duration (absolute value as it doesn't matter whether it was early or late, but what matters is the degree of error).



To create this graph I grouped the previously joined tables by their common IDs and then calculated the delivery time medians. I decided to compare the delivery times by sector to the planned delivery times, however the planned delivery times are equal because of the way that the value is calculated. To solve this, I submit my proposal in the “research.pdf”.

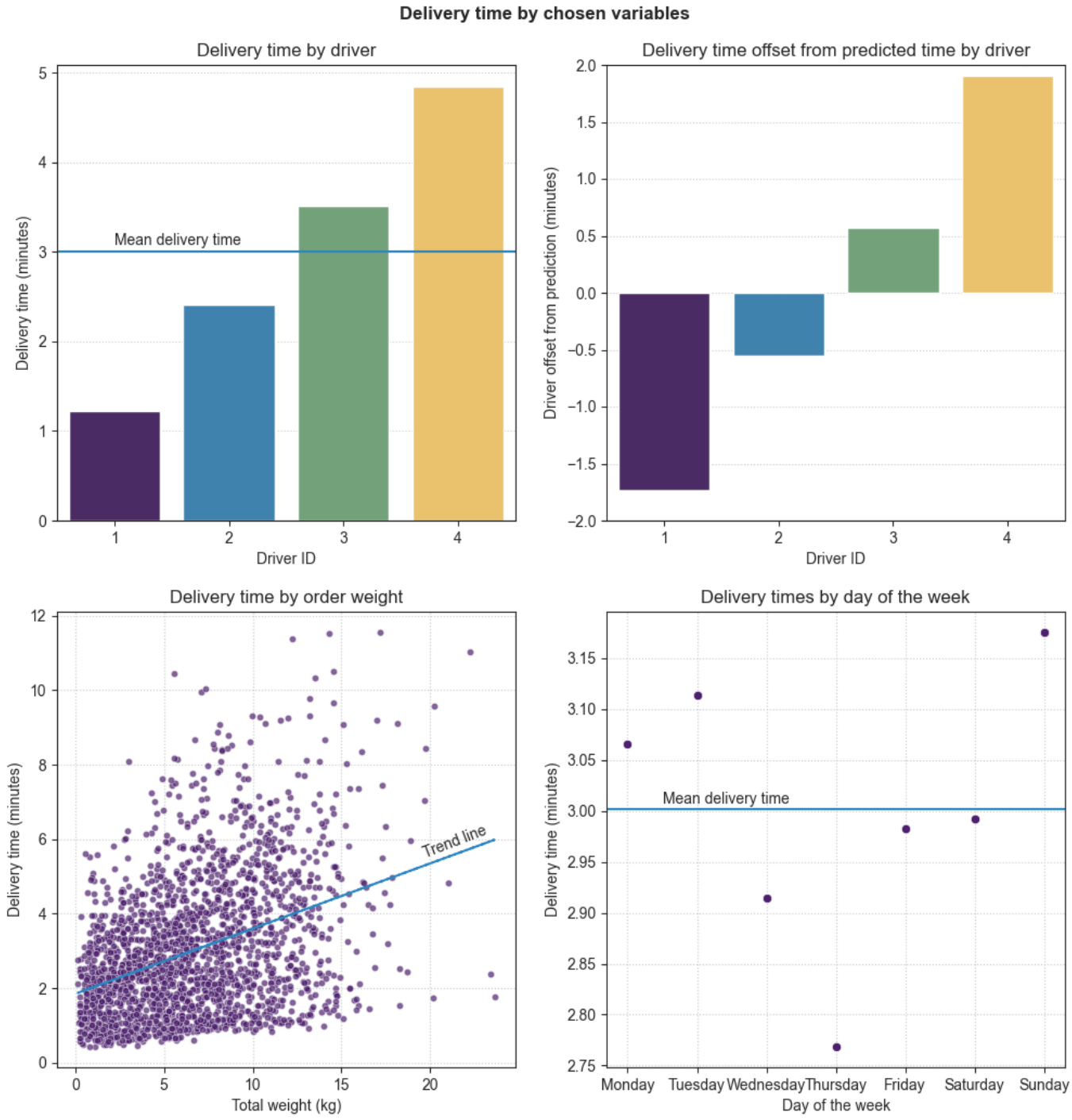


To explore the data and try to find some trends in regards to delivery time I tried multiple approaches.

Chart 1 & 2: I explored the relationship between the order's delivery time and its' driver. First, I calculated average delivery times for each of the drivers. From the charts you can clearly see that delivery time is heavily dependant on the driver. On Chart 2, to quantify how far away each driver is from the expected values, I calculated the offset of their delivery time means to the expected time. Again, as the expected time is the same for each driver the results are basically a different way of portraying Chart 1, however if the expected times were, for example, unique to each sector, Chart 2 would've been more useful.

Chart 3: I explored the relationship between the total weight of the orders and its' delivery time. To get the total weight I merged the tables on their common IDs, multiplied the quantity of the ordered product by weight of a single unit and then summed for each order. Based on Chart 3 and the correlation coefficient between the variables I found that there is a weak-moderate relationship between them i.e. the heavier the package, chances are the longer it will take to deliver.

Chart 4: I explored the relationship between delivery time and the day of the week the order was made. To get the weekday I extracted the date from the route segments table. Then I grouped by the weekday, calculated the mean of the delivery times and visualized it. From Chart 4 you can read that the delivery times get longer at the start and end of the week, but get faster (up to almost 8 % from the mean on Thursday) in the middle of the week.


Delivery time by chosen variables
Delivery time by driver
Mean delivery time
Delivery time (minutes)
Driver ID
Delivery time offset from predicted time by driver
Driver offset from prediction (minutes)
Driver ID
Delivery time by order weight
Trend line
Delivery time (minutes)
Total weight (kg)
Delivery times by day of the week
Mean delivery time
Delivery time (minutes)
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the week