**Capstone Project: analysis of Car accident severity**

Bakhmutova Anastasia

October, 2020

1. **Introduction**
   1. **Background**

There are many car accidents every day and every minute. Each accident brings a strong surge of emotions among the participants. For the most part, these are negative emotions such as fear, sadness, anger. But there are some that give relief.

As a result of such incidents, there may be deterioration in human health, crashes of cars and other vehicles, damage to surrounding objects, and the creation of a traffic jam. And in the worst case, people die. Car accidents can occur for a variety of reasons, such as poor road visibility, poor quality roads, unseen car, etc.

Each case has its own reasons for the incident. It would be quite helpful to analyze them and try to reduce their impact.

* 1. **Problem**

Car accidents have always been terrible moments in people's lives, regardless of age. There are many reasons that can cause a car accident. We will limit ourselves to such external factors as inattention, being under the influence of drugs or alcohol, weather conditions and so on. It is also worth considering where accidents occur most often and mark these places as the most dangerous. And the degree of damage to the participants in the accident will let us understand the severity of the accident. In this way, we will use the data to identify common car accident locations and their causes. We also categorize the reasons into artificial and natural. Accordingly, we will be interested in those data attributes that will help our problem.

* 1. **Interest**

Analysis of data on accidents that have already occurred can help avoid future tragedies. The research results can be useful to both drivers and the state. Drivers will be familiarized with the statistics of collisions in the places where they usually drive and their cause. This will allow drivers to be either more attentive and slower in this section of the road, or use special upgrades for the car. The state will be familiarized with the main causes of accidents, which will help improve the control system on the roads. If the reason was the "absence" of the road, then it is obvious that it is worth building. In other cases, you can install more special lighting against fog or put more police traffic controllers.

1. **Data acquisition and cleaning**
   1. **Data sources**

All the information with which the research was done was taken from the course of ‘Applied Data Science’ at Coursera site. The data are presented in a single table and they were enough for analysis.

* 1. **Data cleaning**

A large table was given, requiring data processing. First, there were 2 columns that described the time of the accident. But in one case there was information only about the date of the accident, in the second case the exact time of the accident was clarified. Therefore, it was decided to leave the attribute INCDATE as it could provide more information.

However, the work with the time column is not complete. After analyzing, it became clear that there is a lack of data for 2019 and 2020. It is believed that there were too few accidents in these years compared to other years to draw any conclusions on their basis. So our project is studied since 2005 to 2018.

Consider columns with coordinates, i.e. X and Y. Since one of the tasks of the research is to find dangerous places where incidents most often occurred, then it is important for us to have information in these columns. The rows of the table, where this data was empty, were deleted.

Two more columns which describe similar things are SEVERITYCODE and SEVERITYDESC. It was decided to leave the description, because this is more useful for clarity. The column EXCEPTRSNCODE was next. This attribute was removed due to the fact that there was a lot of missing data. For the same reason removed the column PEDROWNOTGRNT.

* 1. **Feature selection**

After data cleaning, there were 177,774 samples and 34 features in the data. Since we set goals for ourselves, it turned out that some of the columns would not be useful in the future. Therefore, such columns as INCKEY, COLDETKEY, SDOT\_COLCODE, SDOT\_COLDESC and so on, have been removed.

There were also functions, the study of which would give some results, but within the framework of the tasks set, their study in this project does not cause the need. For example, HITPARKEDCAR, JUNCTIONTYPE, COLLISIONTYPE and ect.

In the end, we used ‘pandas.DataFrame.dropna’ to avoid complications in further work. As a result, we got a table from177,774 samples and 29 features.

1. **Analysis section**
   1. **Location**

First of all, it was decided to identify the most common crash sites: it is logical to assume that these are also the most dangerous places to drive. To do this, we separate the data on the crash sites into a separate Dataframe and group the data, creating another column as 'Amount of Collisions'. To make it easier to understand the results, sorting was done and the data is shown in descending order. It should be noted that we did not take out the places of car accidents, but only 10. In other cases, there were less than 100 accidents.

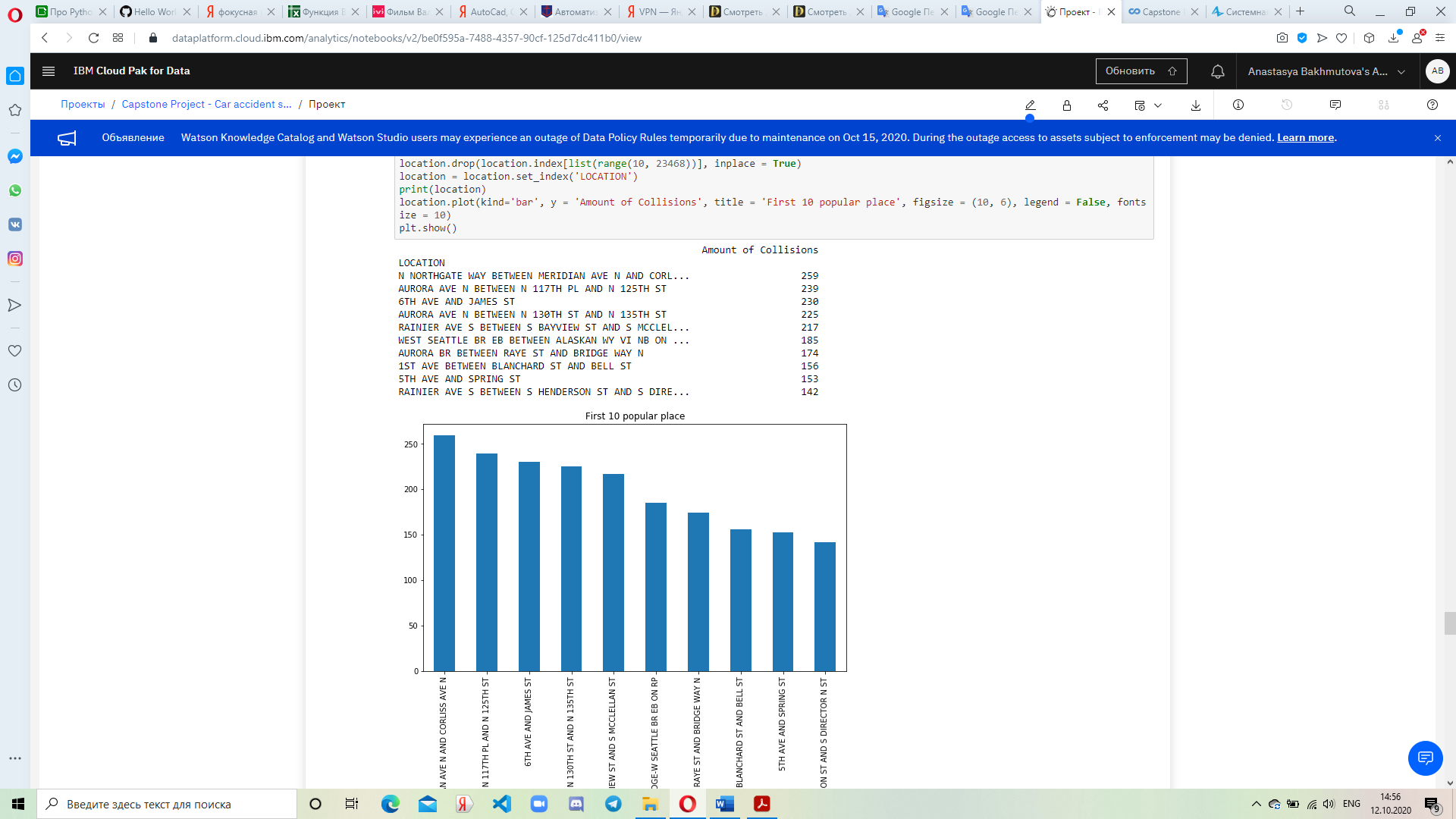


Figure 1. Table of the most popular locations

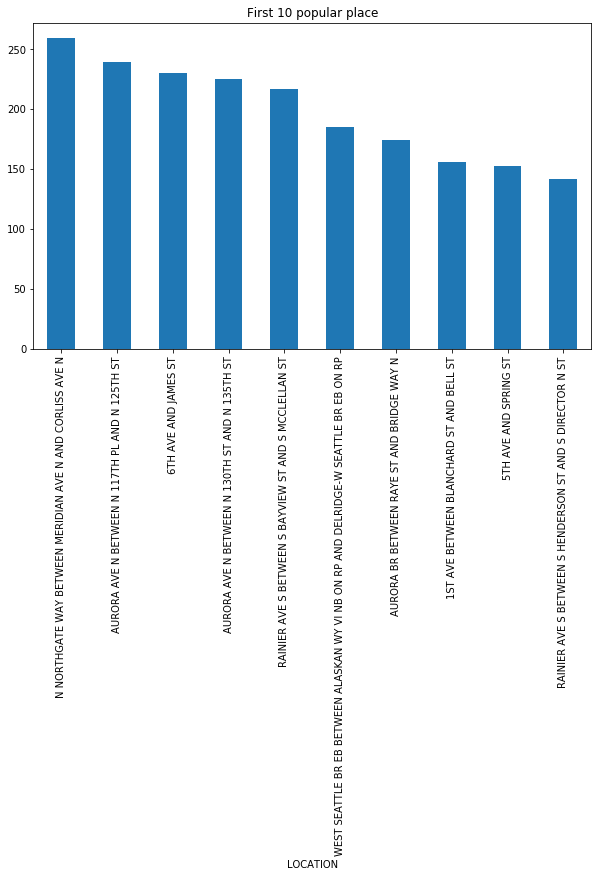


Figure 2. Bar chart of the most popular locations.

In addition to the location of car accidents, we also considered the types of these places, i.e. whether these places are intersections, alleys or blocks. This analysis was accepted because the type of road section may affect driving conditions or the number of people involved in an accident. As a result, the answer was received that blocks are 117 515, intersections are 60 259 but there is no any alley.

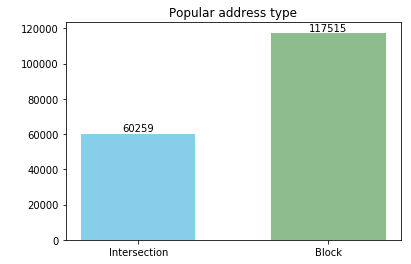


Figure 3. Popular address type

* 1. **Checking time**

Our next step is to consider the timing of accidents. Here we decided to divide all the time into 2 parts. The first group includes the time between the sunrise and sunset, the second group includes the rest. It has been found that in the United States of America the sun usually rises at 8 am and decreases at 6 pm. This assumption seemed to us quite interesting, since the inclusion of additional artificial light depends on the sun.

So, in the course of working with time, we selected data showing the time in hours. With each new indicator of a certain counter, it increased by one. After reviewing all the data, the bar chat was created. Interestingly, most of the accidents happened during daylight hours, when there should be sufficient lighting.

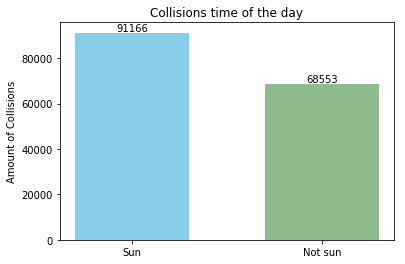


Figure 4. Collisions time of the day

Also, the causes of accidents can be specific weather conditions that belong to a certain season of the year, so we decided to divide accidents into 4 groups according to the number of seasons. Here we also split the dates into components and took only months. At the end, oddly enough, it turned out that most of the accidents did not happen in winter

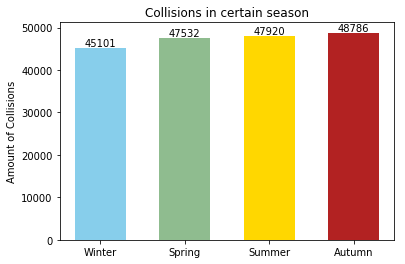


Figure 5. Collisions in certain season.

* 1. **Severity**

The consequences of accidents can be used to assess their severity. For us, the results of accidents are different types of damage: from just damage to property to the death of one of the participants. We made sure that the data for all accidents is known. Moreover, after conducting the analysis, we made sure that absolutely all the crashes occurred without serious injuries to the participants or without their deaths. Only damage to property and minor injuries are observed, while people were injured only in 30% of cases. As a result, we can draw a small conclusion that the accidents that occurred were relatively safe for the health of drivers and passengers.

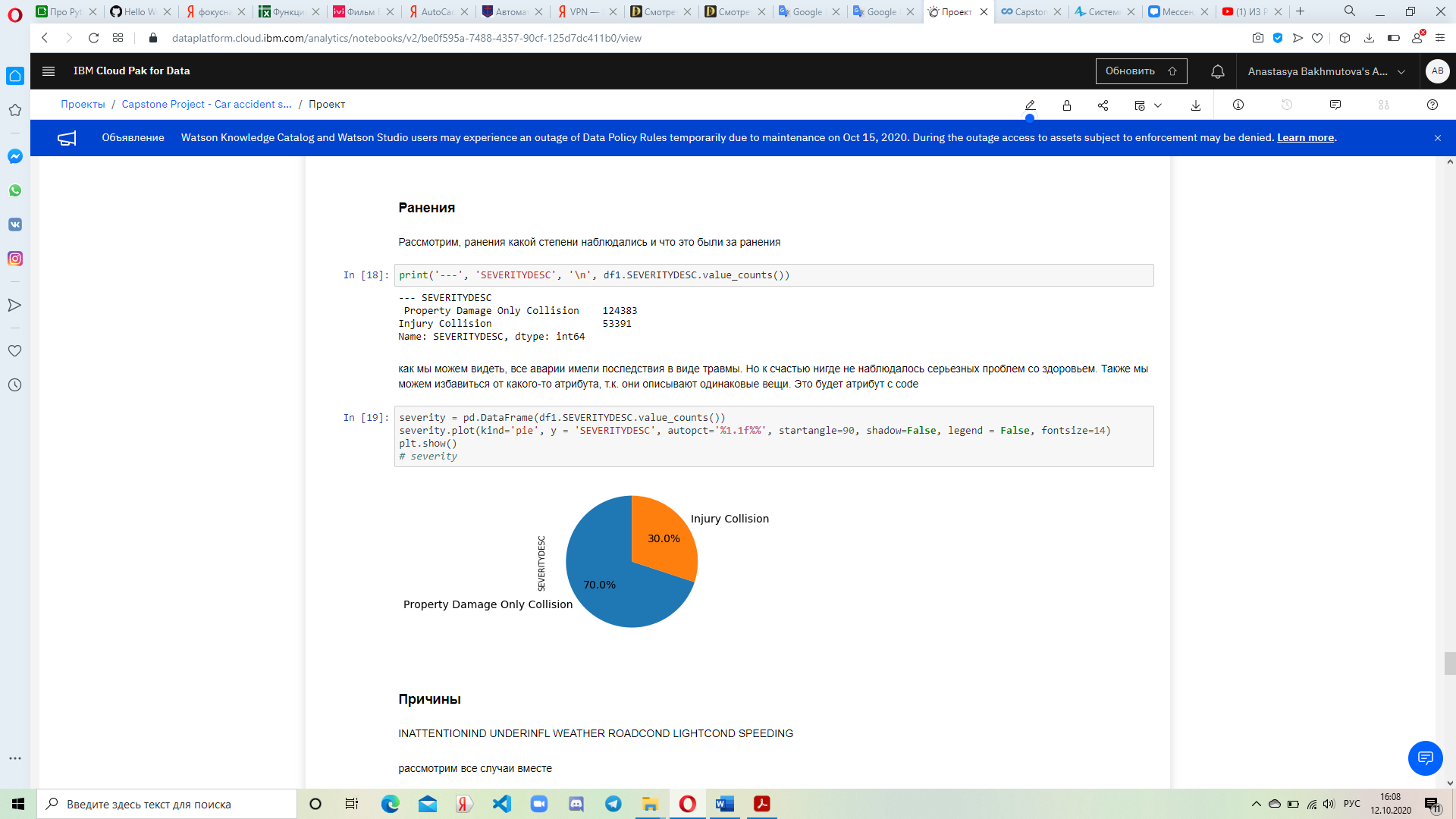


Figure 6. Table of the severity of the collisions

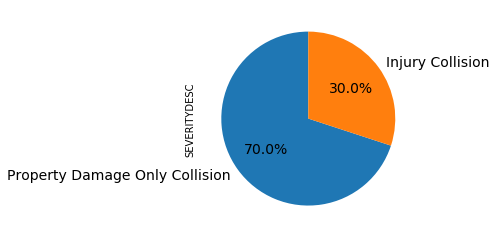


Figure 7. Percentage of the severity

* 1. **Reasons of the collisions**

For a clear understanding of what could have caused the car crash, we will consider all the columns of causes that are given in the original data. In such columns, usually the answers are divided into positive or negative answers, so in certain situations some reasons will be automatically removed from the possible ones.

First of all, we needed to bring the column data into a form convenient for research. For example, in the attribute indicating whether the driver was under the influence of drugs or alcohol, the data values ​​were: 0, 1, Y, or N. We allocated 1 and Y as Yes, 0 and N as No.

Putting the data in order, we performed some manipulations with the possible causes of car accidents and came to the following results:

- For the most part, the drivers were attentive while driving;

- More than 90% of drivers did not use drugs or alcohol;

- The weather was fine and clean;

- There was sufficient bright daylight;

- More than 95% of drivers did not exceed the legal speed.

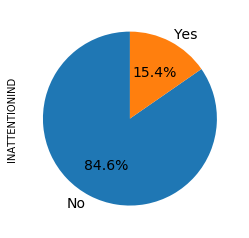
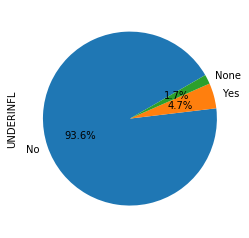
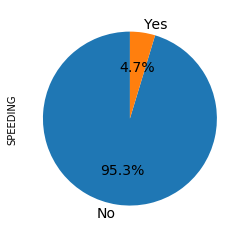
The only negative aspect is the condition of the road. In most cases, the road was dry.

Figure 8. Inattention Figure 9. Influence of drug or alcohol Figure 10. Speeding

It is worth noting that for the conditions of light, roads and weather, we marked the conditions that were encountered most of all, the rest we added to the "Other" category. This is done to make it easier to visualize and understand the data.

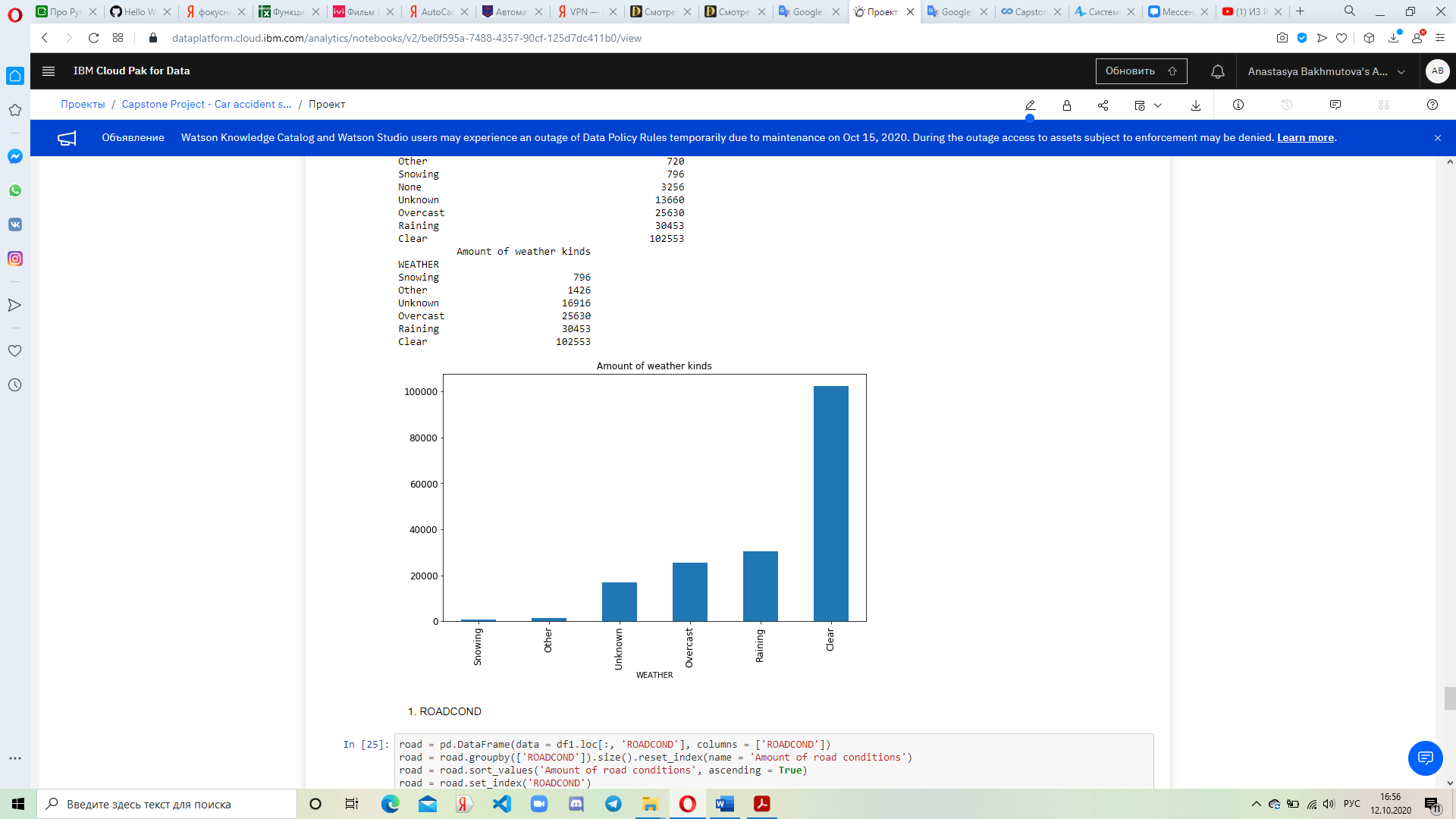


Figure 11. Table of weather conditions

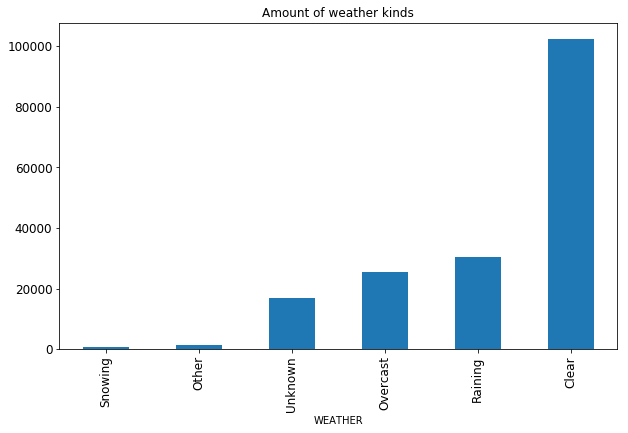


Figure 12. Bar chart of weather conditions

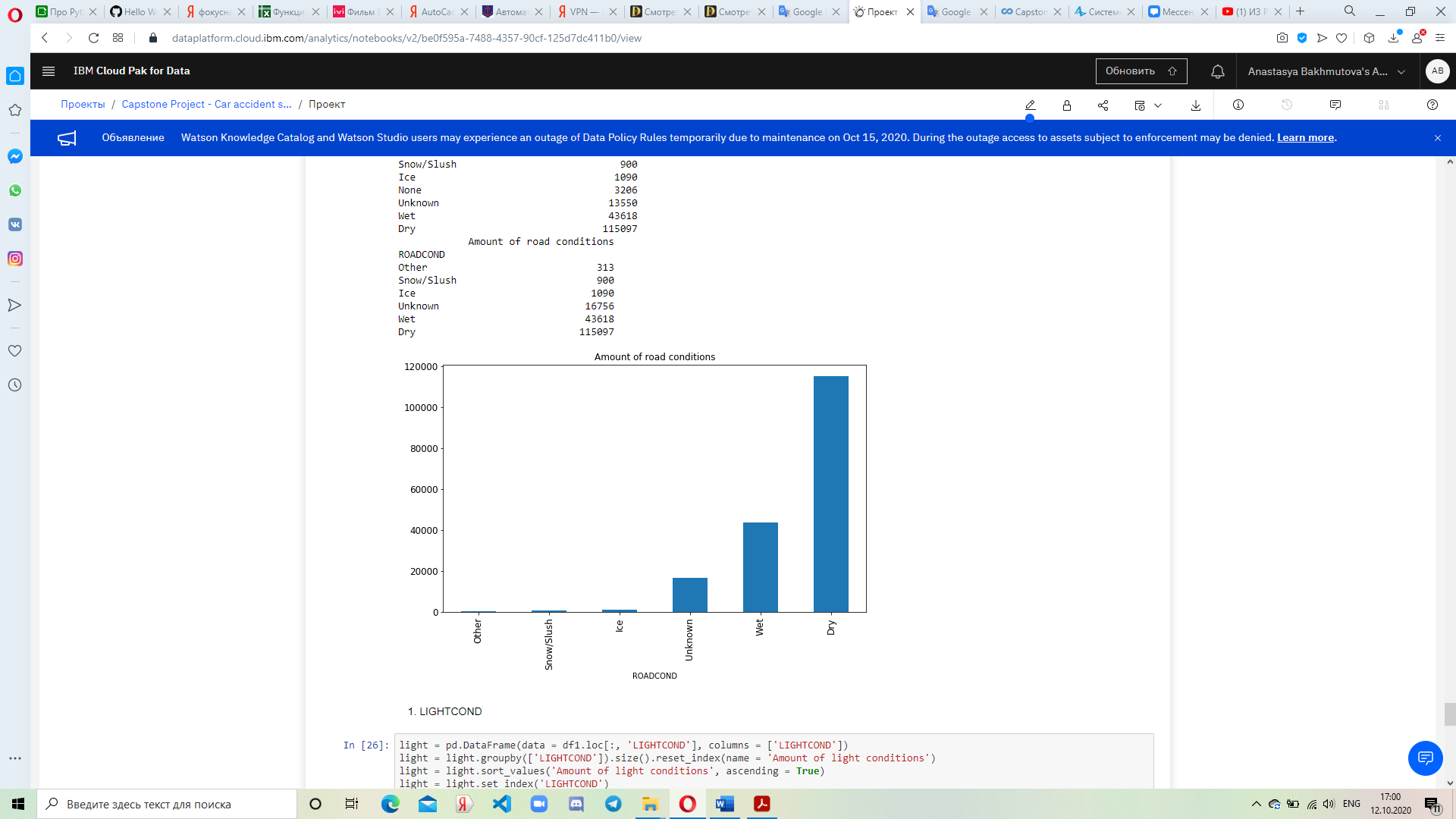


Figure 13. Table of road conditions

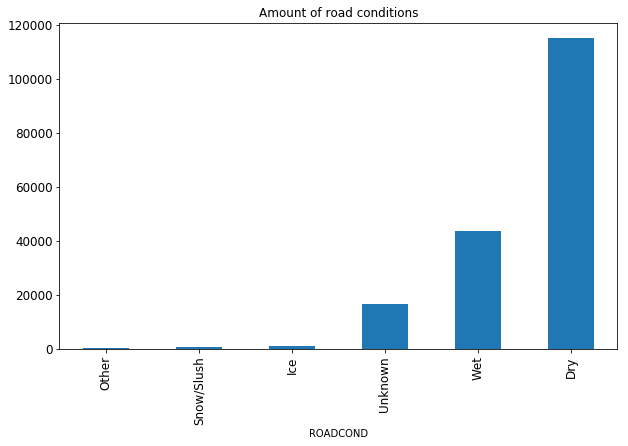


Figure 14. Bar chart of road conditions

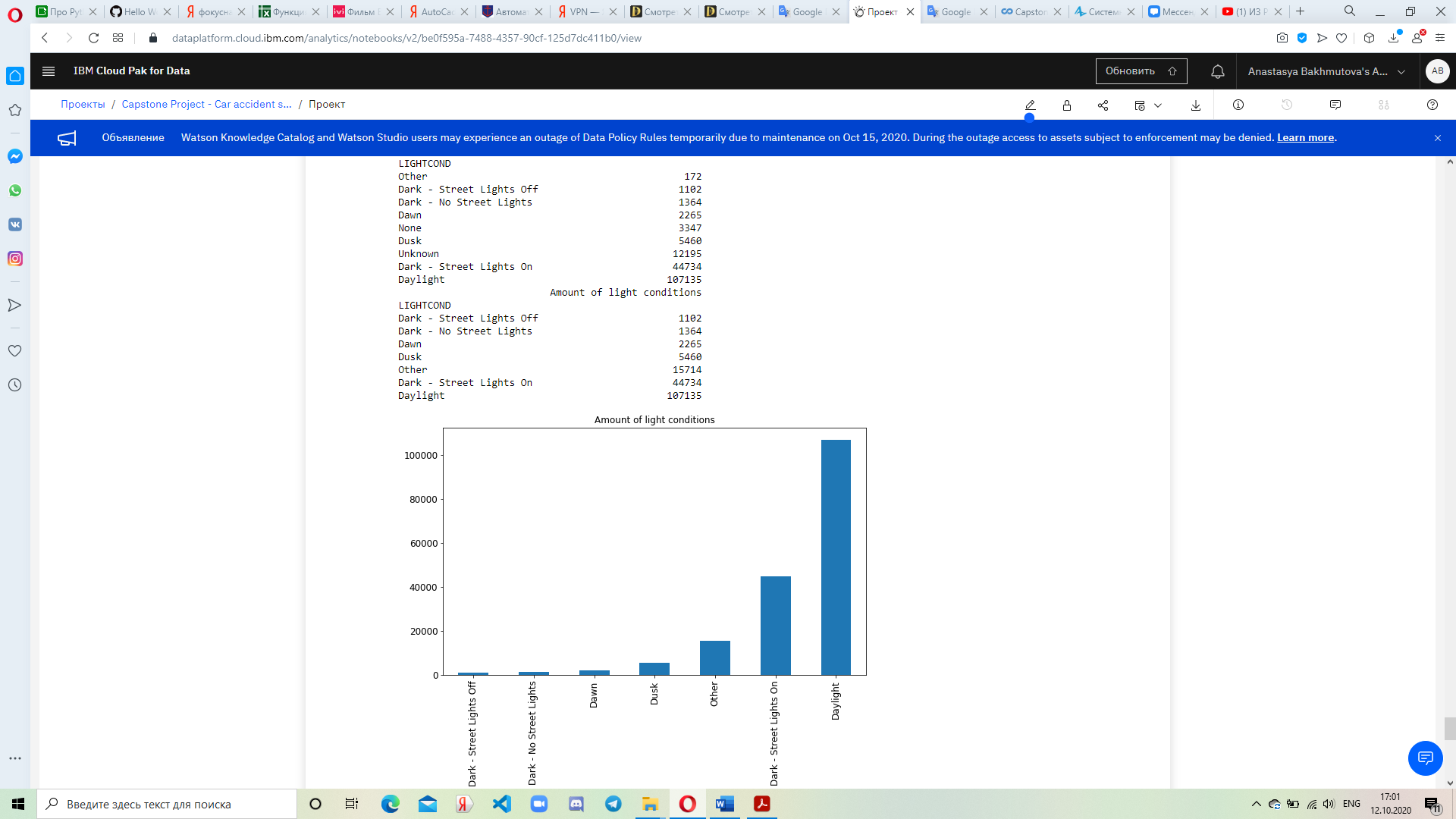


Figure 15. Table of light conditions

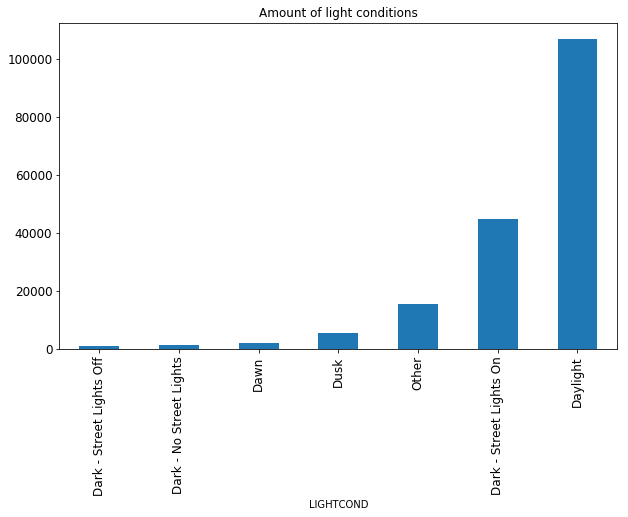


Figure 16. Bar chart of light conditions

* 1. **Car accident participants**

The last thing we noticed was the participants in the incident. This is important because each member of the accident had their own driving rules and characteristics of the movement of their vehicle. For the convenience of presenting information, we have presented data about people using a diagram, and the rest using tables. No special action was required with such data.

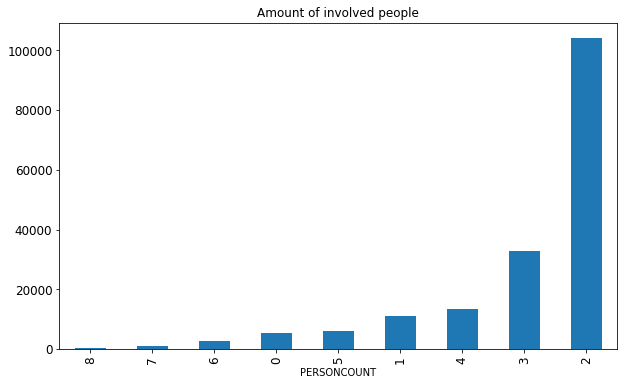


Figure 17. Amount of collisions with certain number of involved people

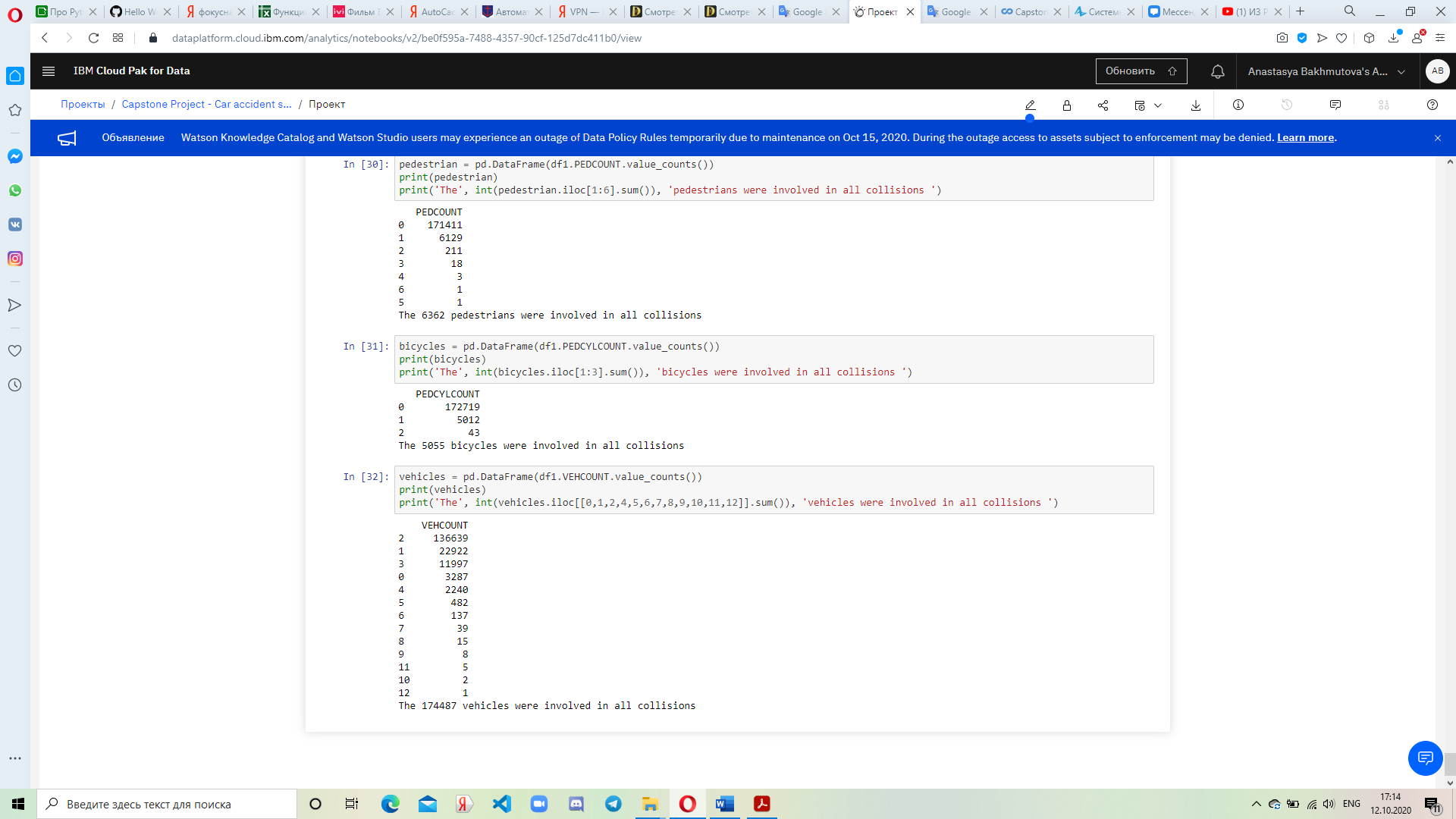


Figure 18. Table of amount of involved pedestrians

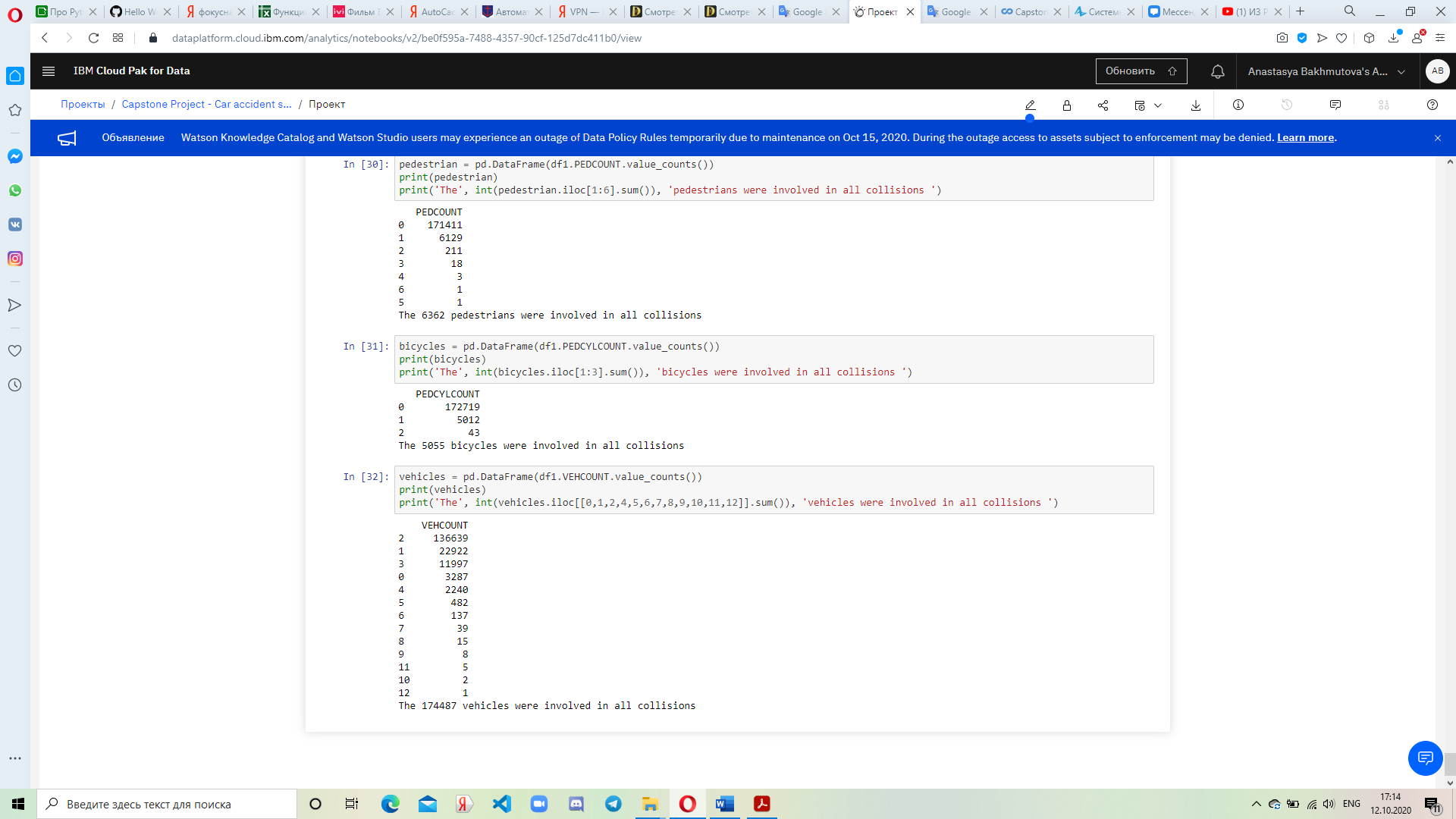


Figure 19. Table of amount of involved bicycles

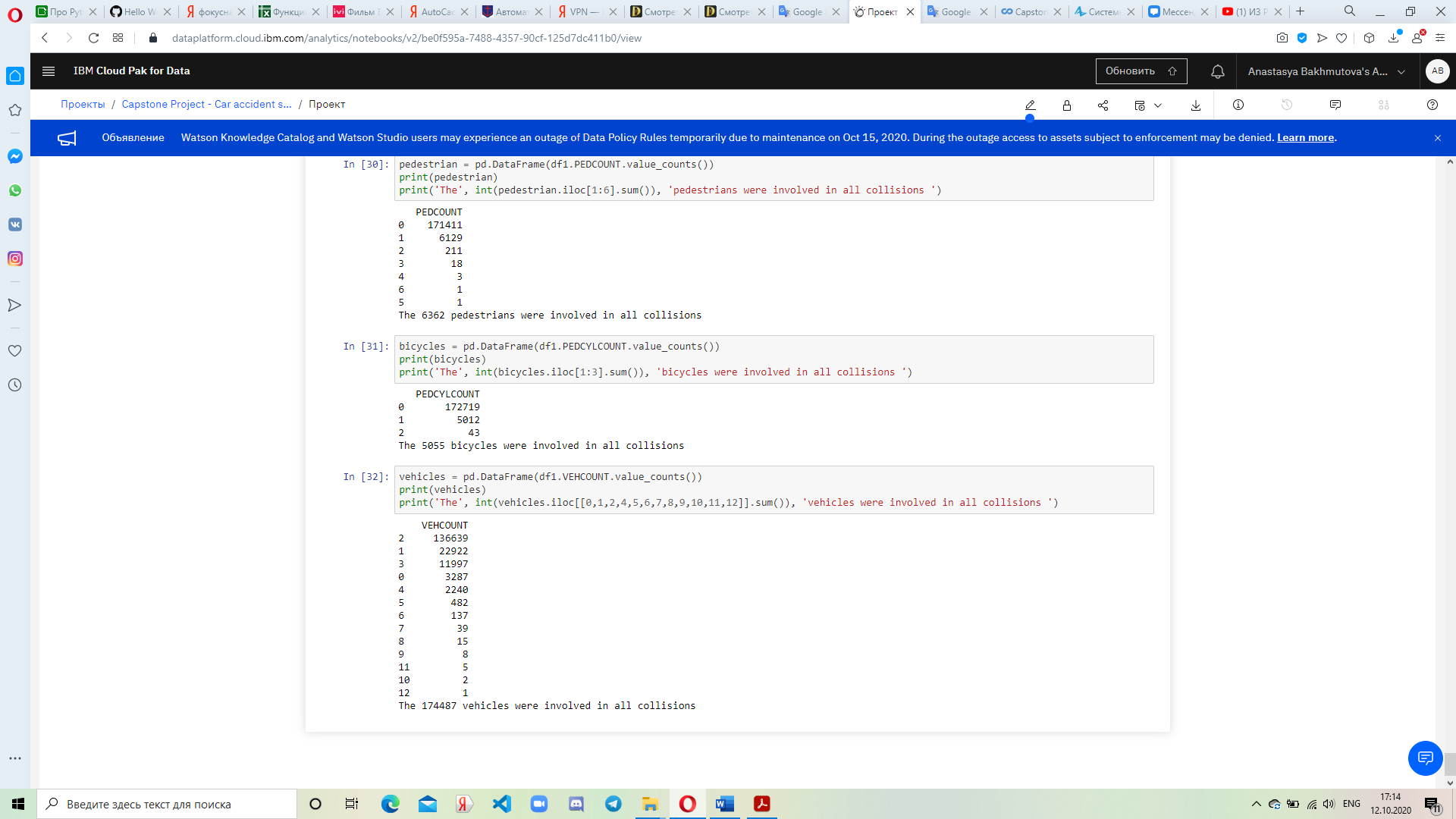


Figure 20. Table of amount of involved vehicles

1. **Results and conclusions**

Summing up the results of the research work, we have the idea that most of the car accidents did not occur due to negative external influences or poor health of the drivers. The only unfavorable quality was the road, because it was wet, which could lead to uncontrolled sliding of the car.

However, all accidents did not have serious consequences for human health. All damage was caused by vehicles or surrounding inanimate objects.

There is also a trend in the number of participants present in the accident. In this way:

- the number of people involved in the accident is usually 2;

- most often 2 cars were involved in accidents;

- cyclists practically did not get into accidents;

- just like cyclists, pedestrians for the most part did not participate in incidents.

Based on such data, it is worth conveying to drivers that these sections of the road are dangerous, so that people become even more attentive. And the state needs to take care of the condition of the roads and, possibly, adopt additional restrictions.

1. **Future directions**

The project has further development prospects. At a minimum, not each cause can be considered separately, but their combined effect on the accident can be studied. In some case, this can give a clear understanding of the fault of the crash. But this stage should be built into a separate project, since takes a long time due to the different combination.

There is also an option to analyze each of the most dangerous places. Identify who the participants in the accidents are, the causes of accidents. This will give a more accurate idea of what to do with this section of the road.