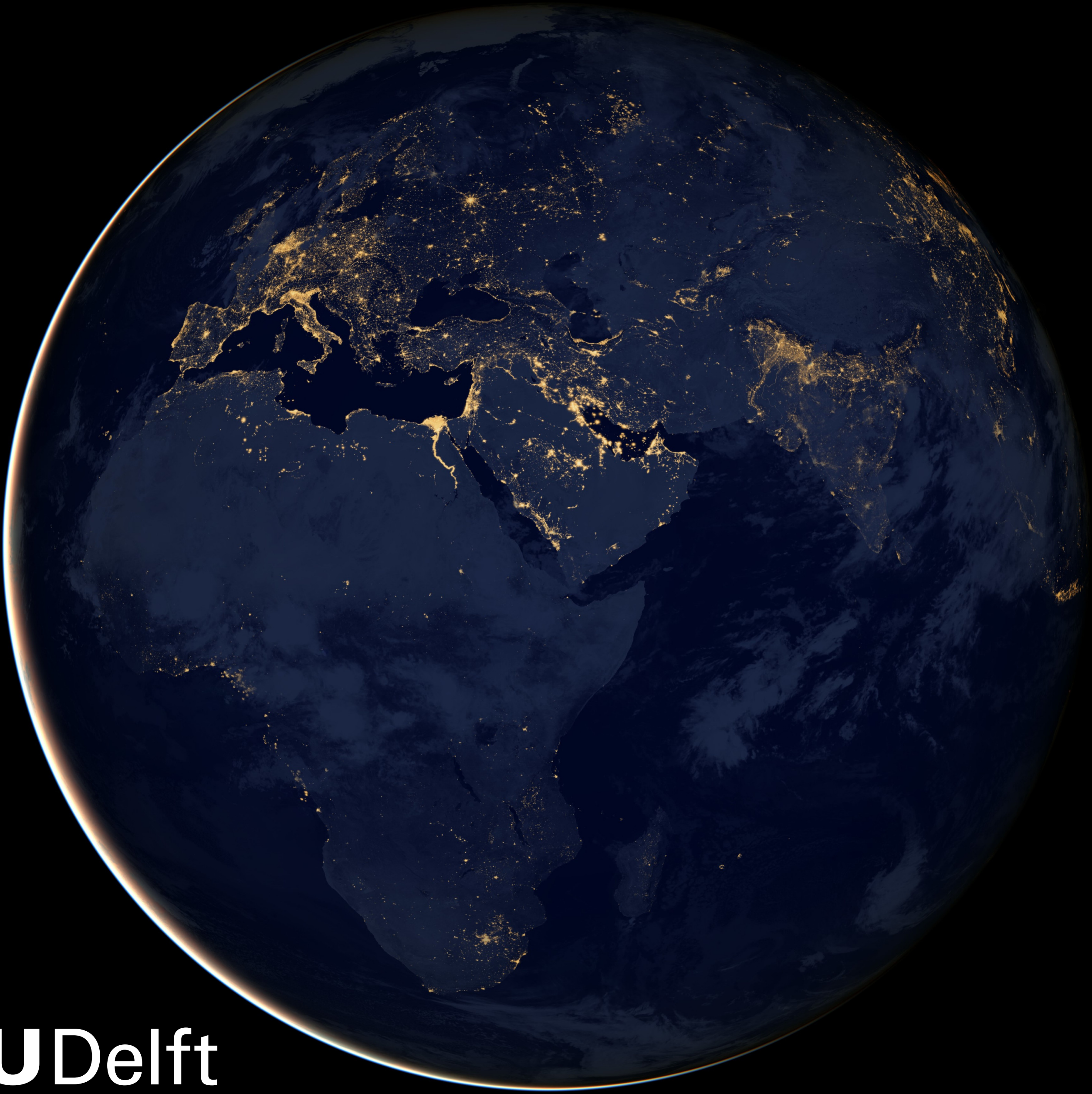
Final Project

Data Visualization and Manipulation

INF 423: Statistics and Visualization for Data Analysis

Dina Kengesbay
Sevda Poladova



Delft University of Technology

TU Delft

Final Project

Data Visualization and Manipulation

by

Dina Kengesbay
Sevda Poladova

Student Name	Student Number
Poladova Sevda	190103364
Kengesbay Dina	190103235

Instructor:	Aigerim Abubakirova
Institution:	Suleyman Demirel University
Place:	Faculty of Engineering and Natural Sciences, Almaty
Project Duration:	one week

Preface

A preface.

Dina Kengesbay
Sevda Poladova
Almaty, December 2021

The purpose of this work is to visualize data for visual representation of the information of announcements with OLX.KZ related to the sale, exchange, recycling of cars[2]. We will explain our data,logical relations of created questions and why we used chosen charts, and how we solved project question: What kind of year might there be in cars depending on their price?

Contents

Preface i

1 Introduction 1

2 Question 2

3 Charts 3

3.1 Chart 1 3

3.2 Chart 2 4

3.3 Chart 3 4

3.4 Chart 4 5

3.5 Chart 5 5

3.6 Chart 6 6

3.7 Chart 7 6

3.8 Chart 8 7

4 Conclusion 8

5 References 9

1

Introduction

We started working with a database that is associated with announcements about cars from the site OLX.KZ[2]. Our dataset originally consisted of five columns:

1. Unnamed: 0 - indexes
2. Title - announcement
3. Price - car price
4. City - the city where the car and its seller are located,publication date
5. Link - link to the original announcement on the website - OLX.KZ.

First, we cleaned our data set of unnecessary columns, such as ”Unnamed: 0” and ”Link”, because they are not important for visualization. We have also cleared our dataset of duplicates. Then, using various data manipulations, we were able to expand our dataset.After analyzing the column ”Title”, we came to the conclusion that some announcements are urgent, and some are about the sale, exchange and recycling of cars. Also analyzing the column ”City”, we came to the conclusion that in order to use this column in visualization, we need to divide it into two columns - the city and the day of publication. We also tried to identify the marks of cars from the column ”Title”.Thus, our data set has become much larger and it began to consist of 9 columns:

1. Title - announcement
2. Price - car price
3. City - the city where the car and its seller are located
4. date - publication date
5. isUrgently - urgent announcements
6. isRecycling - announcements about the disposal of cars
7. isExchange - announcements on the exchange of cars
8. isSell - announcements for the sale of cars
9. carmarks - mark of car

2
Question

Question: What kind of year might there be in cars depending on their price? Hint: Categorize the price range and year range, then use these parameters to identify the year.
The first thing we did in this question was categorize the price range. We took the interval between the maximum and minimum price and divided it into five equal intervals.In order to do this, we wrote the following code:
a = np.linspace(data[’Price’].min(), data[’Price’].max(), 6)

This code return the following array:
array([77777.,3362221.6, 6646666.2,9931110.8,13215555.4,16500000.])

Using this 6 numbers, we created 5 intervals.Based on these intervals, we create a new column ”Category”, which consists of the following classifications: very cheap, cheap, middle, less expensive, expensive, unknown. Code how we create a new column based on the obtained intervals:

```
conditions3 = [
    (data[’Price’] >= a[0]) & (data[’Price’] <= a[1]),
    (data[’Price’] > a[1]) & (data[’Price’] <= a[2]),
    (data[’Price’] > a[2]) & (data[’Price’] <= a[3]),
    (data[’Price’] > a[3]) & (data[’Price’] <= a[4]),
    (data[’Price’] > a[4]) & (data[’Price’] <= a[5]),
    (pd.isna(data[’Price’]))
    ]
values3 = [’Very□cheap’, ’Cheap’, ’Middle’, ’Less□expensive’,’Expensive’,’unknown’]
data[’Category’] = np.select(conditions3, values3)
```

Then we categorized year range. From 1986 to 2021. We had many old cars, so we started with 1986 year. Almost all intervals with 8 years, expect 2018-2021. Using these two parameters, we identified the year.Code:

```
conditions4 = [
    (data[’Category’] == ’Very□cheap’),(data[’Category’] == ’Cheap’),
    (data[’Category’] == ’Middle’),(data[’Category’] == ’Less□expensive’),
    (data[’Category’] == ’Expensive’),(data[’Category’] == ’unknown’)
    ]
values4 = [’1986−1993’, ’1994−2001’, ’2002−2009’, ’2010−2017’,’2018−2021’,’unknown’]
data[’Year’] = np.select(conditions4, values4)
```

3
Charts

3.1. Chart 1

We used a scatter plot to show the maximum car price of each of the car marks.The size of each circle is the maximum price of each mark. Thus, using the scatter plot, we can see the most expensive marks and the most expensive cars in each mark, not only on the y-axis, but also by their size. For example, we can see that the maximum price of Lada from all Ladas that have been posted on the ad site is 6500000. We also worked with the style, changed the font size, turned on the "Color by Category" option in order to make the chart much more attractive and understandable, etc

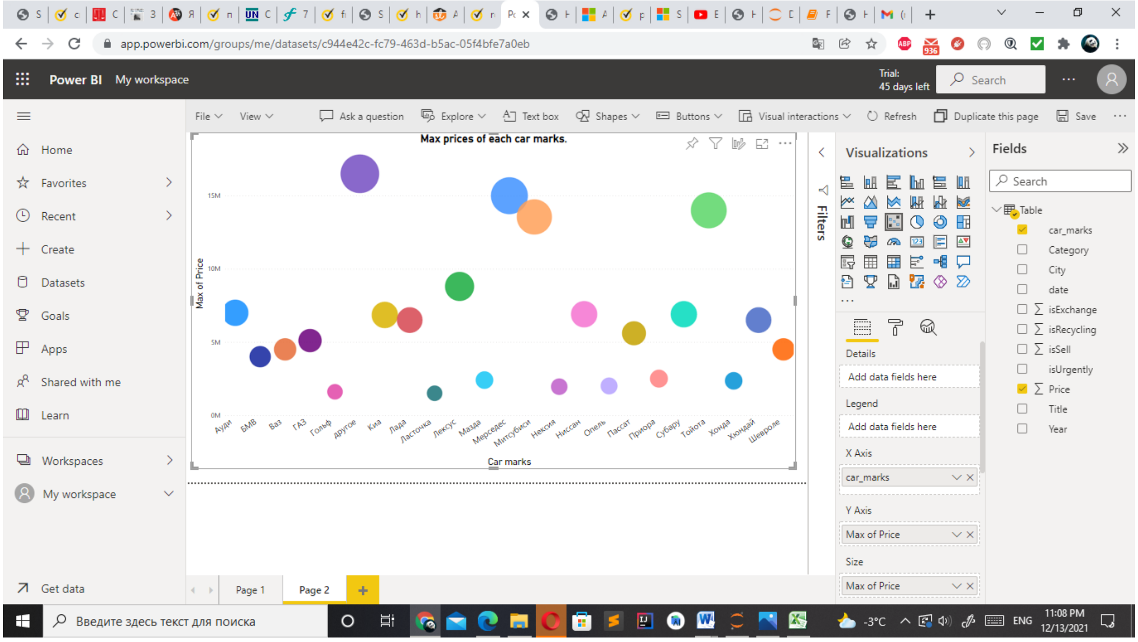


Figure 3.1: Scatter chart

3.2. Chart 2

As we said earlier, analyzing the column ”Title” we came to the conclusion that some people sell, some exchange, and some recycling of cars. Using the clustered column chart, we wanted to show the number of each of the named groups. To construct this chart we used columns: IsSell,IsExchange and IsRecycling. Based on the chart, we can say that the ads of those who sell (356) prevail over those who exchange (21) and recycle (10).

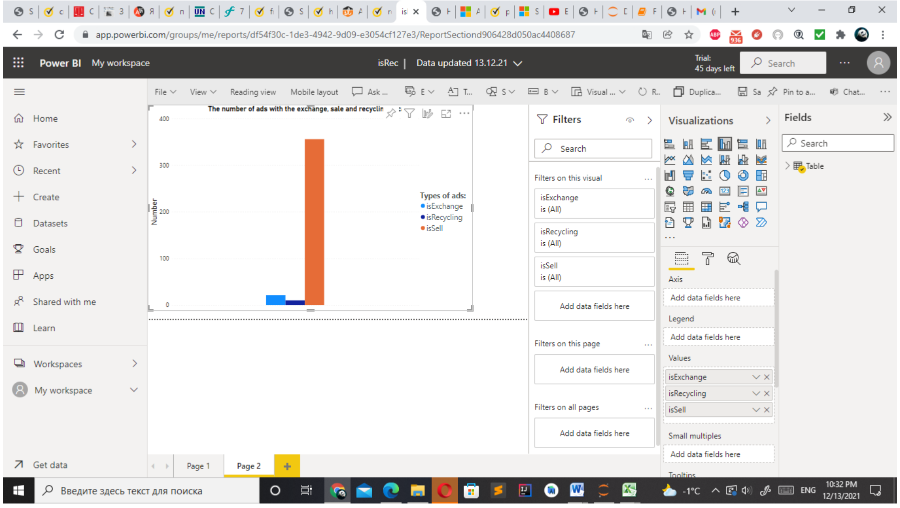


Figure 3.2: Clustered column chart

3.3. Chart 3

We use the pie chart to show the number of ads on a certain date.To construct this chart we used columns: date and Title. Analyzing this chart, we can conclude that there are more and more ads every day. For example, 5 announcements were made for ”April”, 69 for ”May”, 254 for ”yesterday” and 332 for ”today”.

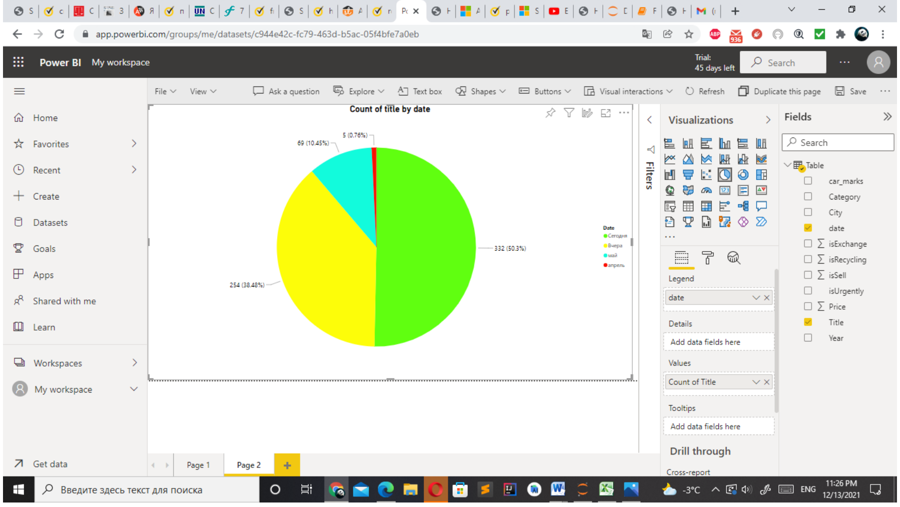
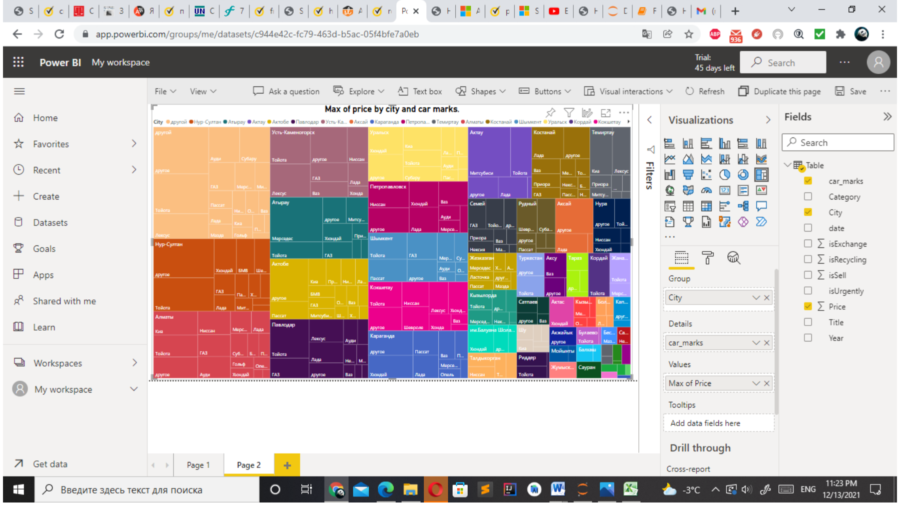


Figure 3.3: Pie chart

3.4. Chart 4

Using the treemap, we wanted to show the maximum price by city and car mark. This means that from each city we will see all the marks of cars that are in this city with the maximum price of one of the cars of this mark in this city.To construct this treemap we used columns: City, car marks and price(with filter max).



app.powerbi.com/groups/me/datasets/c944e42c-fc79-463d-b5ac-05f4bfe7a0eb

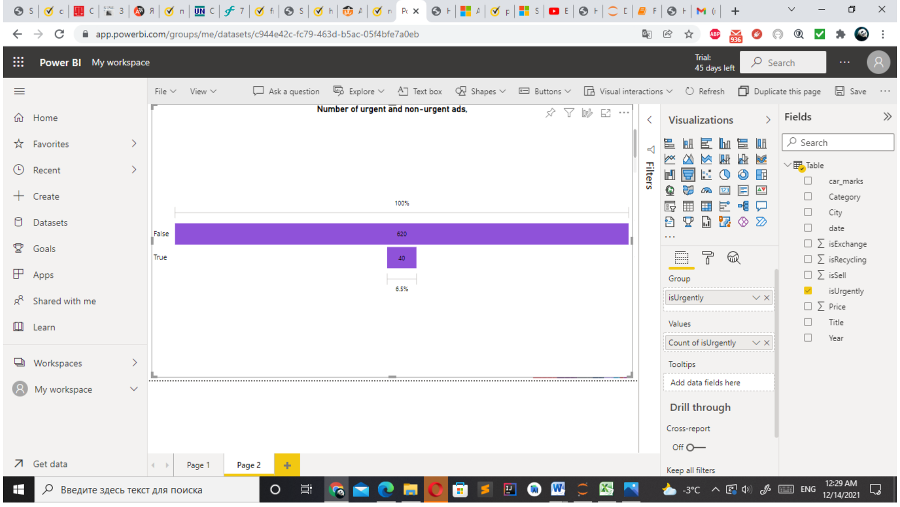
Max of price by city and car marks.

Group: City
Details: car_marks
Values: Max of Price

Figure 3.4: Treemap

3.5. Chart 5

We used funnel chart in order to illustrate amount of urgent and non-urgent announcements. We took the column which called 'isUrgent' which contains values as 'True' for urgent ads and 'False' for non-urgent ads. It can be clearly seen that rate of 'True' values(urgent ads) are less than rate of 'False' values(non-urgent) ads. Urgent ads make up only 6.5(40) percent of the total, while non-urgent ads make up the remaining 93.5(620) percent. We also chose purple for the graphics style and changed the fonts etc.



app.powerbi.com/groups/me/datasets/c944e42c-fc79-463d-b5ac-05f4bfe7a0eb

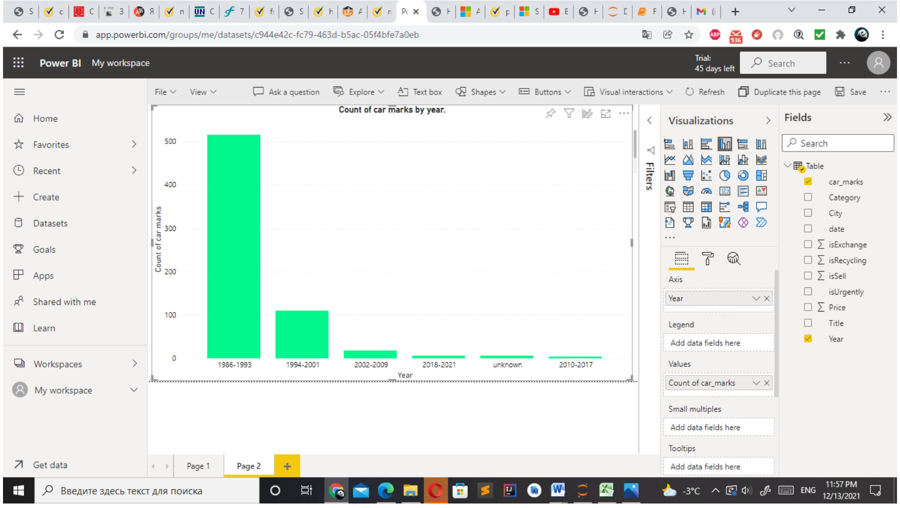
Number of urgent and non-urgent ads.

Group: isUrgently
Values: Count of isUrgently

Figure 3.5: Funnel chart

3.6. Chart 6

Given bar chart displays count of car marks in each year in descending order. As shown in the graph about 500 car brands belong to 1986-1993, approximately 100 to 1994-2001 and only a small number of car marks belong to 2002-2009, 2010-2017, 2018-2021, and the rest are not known for us. Now through this chat we know that most of all car brands belonged to 1986-1993, and least of all to 2010-2017. This time we chose green colour for data in chart.



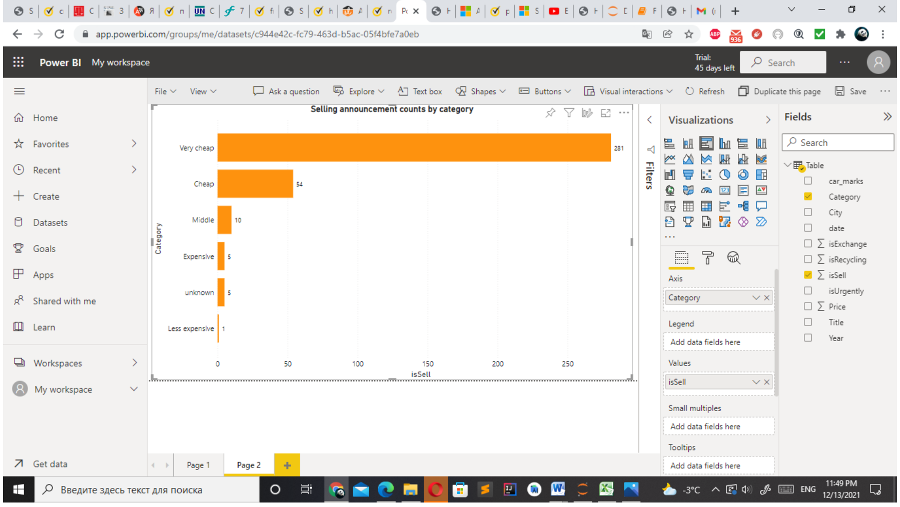
Axis: Year
Legend: Add data fields here
Values: Count of car_marks
Small multiples: Add data fields here
Tooltips: Add data fields here

Table
☑ car_marks
☐ Category
☐ City
☐ date
☐ Σ isExchange
☐ Σ isRecycling
☐ Σ isSell
☐ isUrgently
☐ Σ Price
☐ Title
☑ Year

Figure 3.6: Bar chart

3.7. Chart 7

Following chart known as normalized bar chart and demonstrates amount of announcements regarding to the sale by categories like very cheap, cheap, middle, less expensive, expensive, unknown by descending order of count. According to this ratio, in the graph we see that most of all are sold very cheap cars, and in second place are cheap cars, and less expensive and expensive cars are rarely sold. We marked the data in orange colour for the style.



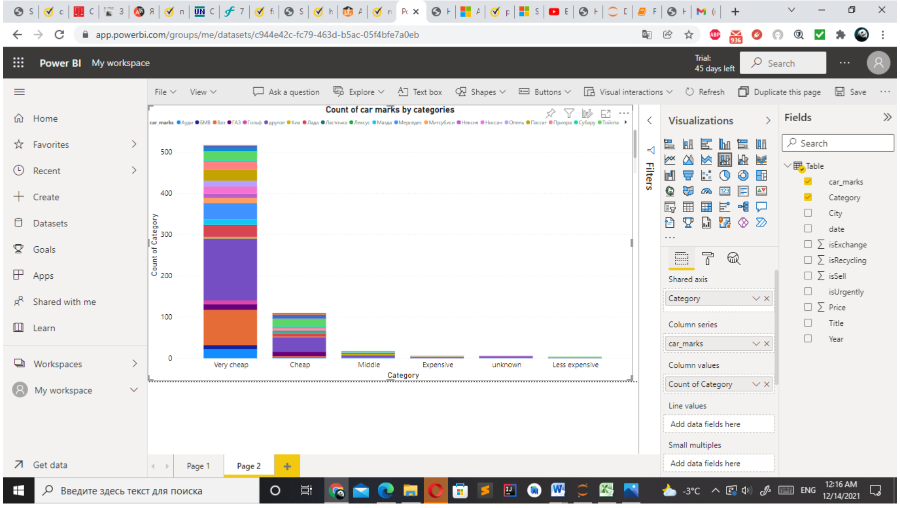
Axis: Category
Legend: Add data fields here
Values: isSell
Small multiples: Add data fields here
Tooltips: Add data fields here

Table
☐ car_marks
☑ Category
☐ City
☐ date
☐ Σ isExchange
☐ Σ isRecycling
☑ Σ isSell
☐ isUrgently
☐ Σ Price
☐ Title
☐ Year

Figure 3.7: Normalized bar chart

3.8. Chart 8

Our last chart is normalized histogram which indicates amount of car marks by categories in decreasing order. At the bottom in the x-axis the categories are indicated, and on the y-axis the number of each car brand. each brand is marked with a different color. For instance, the blue color belongs to mark Audi, purple is BMW, red means VAZ and so on.



Shared axis: Category

Column series: car_marks

Column values: Count of Category

Figure 3.8: Normalized histogram

4

Conclusion

We worked with the cars dataset and visualized it by creating 8 charts in which each of the graphs answers a certain question that we have inventedб also provided a mobile version. We have implemented visualization in Power Bi using different types of charts.

First chart answers the question that how much the most expensive car costs for each car mark.

Second chart answers the question that how many cars need to be exchanged how many to sell and how many to recycle.

Third chart answers the question that how many ads are posted in OLX every day. According to the our dataset we only have the dates like today, yesterday, May and April.

Fourth chart answers the question that how much does the most expensive car in each city cost by mark.

Fifth chart answers the question that how many of the announcements are urgent and how many are not urgent.

Sixth chart answers the question that in what year how many marks of cars.

Seventh chart answers the question that how many cars are sold in each category.

Eighth chart answers the question that how many car marks in each category.

The main goal of our project is to answer the project question what kind of year might there be in cars depending on their price? And categorize the price range and year range, then use these parameters to identify the year. Initially our cars dataset which contains 'Title', 'Price', 'City', 'Link' columns. 'Title' column consists of advertisements of people who want to sell, recycle or exchange their cars urgently. In 'Title' column also was indicated years. As the solution of this question we categorized the price range and year range.

We really enjoyed working with this project. We were interested in its columns and data. We created new columns we need and filled them with values, visualized data and analyzed the changes over the years.

5

References

[1]http://www.overleaf.com

[2]https://drive.google.com/file/d/1D7PB1xsMrvYrG7hyDZLzn_Oq6c3w_m8f/view?usp=sharing