National Research University «Higher School of Economics»

Coursework On Discipline

« Modern Methods of Data Analysis»

Team: Anton Lebedev, Veronica Sarkisyan

Mentor: Boris G. Mirkin

Moscow, 2018

Task 1

Description of the dataset

This dataset was provided by IBM and contains data about telecommunications customers. The aim of exploring this data is to understand why clients decide to refuse operator services and based on the results of the research improve their user experience.

Each row in the dataset represents a customer, each column contains customer's attributes. The dataset includes information about:

- Customers who left within the last month – the column is called *Churn.* Customer churnб also known as customer turnover, or customer defection, is the loss of clients or customers.
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges.
- Demographic info about customers – gender, age range, and if they have partners and dependents.

The dataset consists of 7043 objects (customers), each described by 21 features:

Feature name	Feature type	Description
CustomerID	String	Customer ID
Gender	Categorical	Male / Female
SeniorCitizen	Categorical	Whether the customer is a senior citizen or not (1, 0)
Partner	Categorical	Whether the customer has a partner or not (Yes, No)

Dependents	Categorical	Whether the customer has dependents or not (Yes, No)
Tenure	Numeric	Number of months the customer has stayed with the company
PhoneService	Categorical	Whether the customer has a phone service or not (Yes, No)
MultipleLines	Categorical	Whether the customer has multiple lines or not (Yes, No, No phone service)
InternetService	Categorical	Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity	Categorical	Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup	Categorical	Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection	Categorical	Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport	Categorical	Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV	Categorical	Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies	Categorical	Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract	Categorical	The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling	Categorical	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	Categorical	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
MonthlyCharges	Numeric	The amount charged to the customer monthly

TotalCharges	Numeric	The total amount charged to the customer
Churn	Categorical	Whether the customer churned or not (Yes or No)

There are no missing values in any features. The value we are trying to predict is contained in feature “Churn” that has two entities: “Yes” (if customer churned) and “No” (otherwise).

Task 2

We will use the following features:

- TechSupport
- OnlineSecurity
- InternetService
- DeviceProtection
- PhoneService

Alltogether they describe which services the client uses.

K = 5

labels_5		TechSupport	OnlineSecurity	InternetService	DeviceProtection	PhoneService	labels_5
0	0	No	Yes	Fiber optic	No	Yes	0
1	0	No Internet service	No Internet service	No Internet service	No Internet service	Yes	1
2	0	Yes	No	DSL	No	Yes	2
3	0	Yes	Yes	DSL	Yes	Yes	3
4	0	Yes	No	DSL	Yes	Yes	4

We can observe the following clusters:

Cluster 0: Self-Confidence customers who prefer not to use any additional services: no TechSupport, no DeviceProtection, no OnlineSecurity. They also tend to use Fiber Optic Internet Provider instead of outdated DSL.

Cluster 1: Customers with no Internet at all.

Cluster 2: Opposite to cluster 0 - customers that have all types of additional services: TechSupport, OnlineSecurity. They are more likely to use DSL.

Clusters 3 and 4: Customers with DSL who use one of two services: either OnlineSecurity or DeviceProtection.

K = 9

	labels_9	TechSupport	OnlineSecurity	InternetService	DeviceProtection	PhoneService	labels_9
0	0	Yes	No	Fiber optic	Yes	Yes	0
1	0	No Internet service	No Internet service	No Internet service	No Internet service	Yes	1
2	0	Yes	Yes	DSL	No	Yes	2
3	0	No	No	Fiber optic	No	Yes	3
4	0	Yes	No	DSL	No	Yes	4
5	0	Yes	Yes	DSL	Yes	Yes	5
6	0	No	No	Fiber optic	Yes	Yes	6
7	0	No	Yes	Fiber optic	Yes	Yes	7
8	0	No	Yes	DSL	No	Yes	8

Here we can observe the following interesting clusters:

Cluster 1: Customers with no Internet at all.

Cluster 3: Self-Confidence customers who prefer not to use any additional services: no TechSupport, no DeviceProtection, no OnlineSecurity. They also tend to use Fiber Optic Internet Provider instead of outdated DSL.

Cluster 6: Users like cluster 3, but with DeviceProtection.

Cluster 5: Customers that have all types of additional services: TechSupport, DeviceProtection, OnlineSecurity. They are more likely to use DSL.

Task 3

We will consider the following features:

- tenure
- Churn
- SeniorCitizen

Transform numerical feature “tenure” to categorical:

```
#Tenure to categorical column
def tenure_lab(telcom) :

    if telcom["tenure"] <= 12 :
        return "Tenure_0-12"
    elif (telcom["tenure"] > 12) & (telcom["tenure"] <= 24 ):
        return "Tenure_12-24"
    elif (telcom["tenure"] > 24) & (telcom["tenure"] <= 48) :
        return "Tenure_24-48"
    elif (telcom["tenure"] > 48) & (telcom["tenure"] <= 60) :
        return "Tenure_48-60"
    elif telcom["tenure"] > 60 :
        return "Tenure_gt_60"
telcom["tenure_group"] = telcom.apply(lambda telcom:tenure_lab(telcom),
                                      axis = 1)
```

Contingency tables:

tenure_group	Tenure_0-12	Tenure_12-24	Tenure_24-48	Tenure_48-60	Tenure_gt_60
Churn					
No	1149	730	1269	712	1314
Yes	1037	294	325	120	93

SeniorCitizen	0	1
Churn		
No	4508	666
Yes	1393	476

Quetelet relative index:

tenure_group	Tenure_0-12	Tenure_12-24	Tenure_24-48	Tenure_48-60	Tenure_gt_60
Churn					
No	-0.284514	-0.029592	0.083689	0.164898	0.271255
Yes	0.787627	0.081922	-0.231678	-0.456492	-0.750921

SeniorCitizen	0	1
Churn		
No	0.039895	-0.206148
Yes	-0.110443	0.570686

On average, knowledge of tenure adds 12.3% to the frequncy of Churn, knowledge of SeniorCitizen adds 2.2% to the frequncy of Churn.

For 95% confidence level we need 148 observations.
For 95% confidence level we need 188 observations.

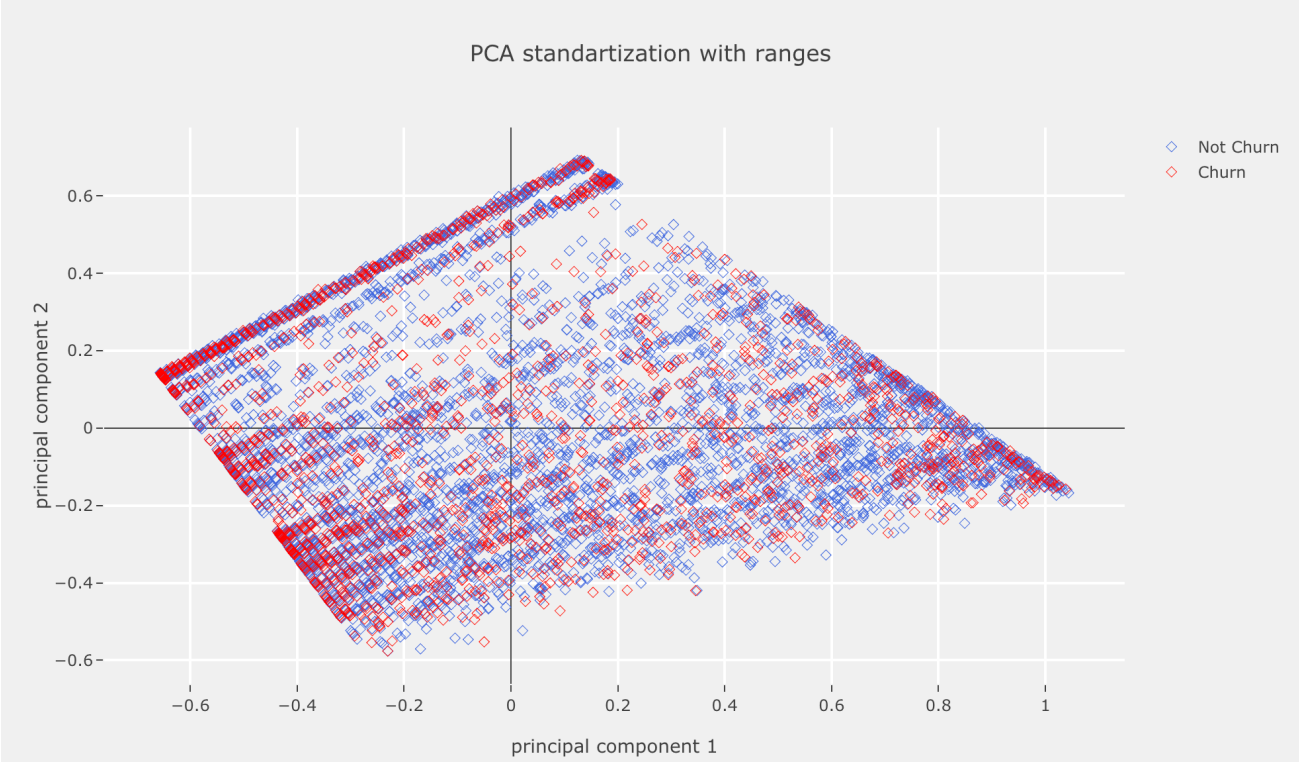
Task 4

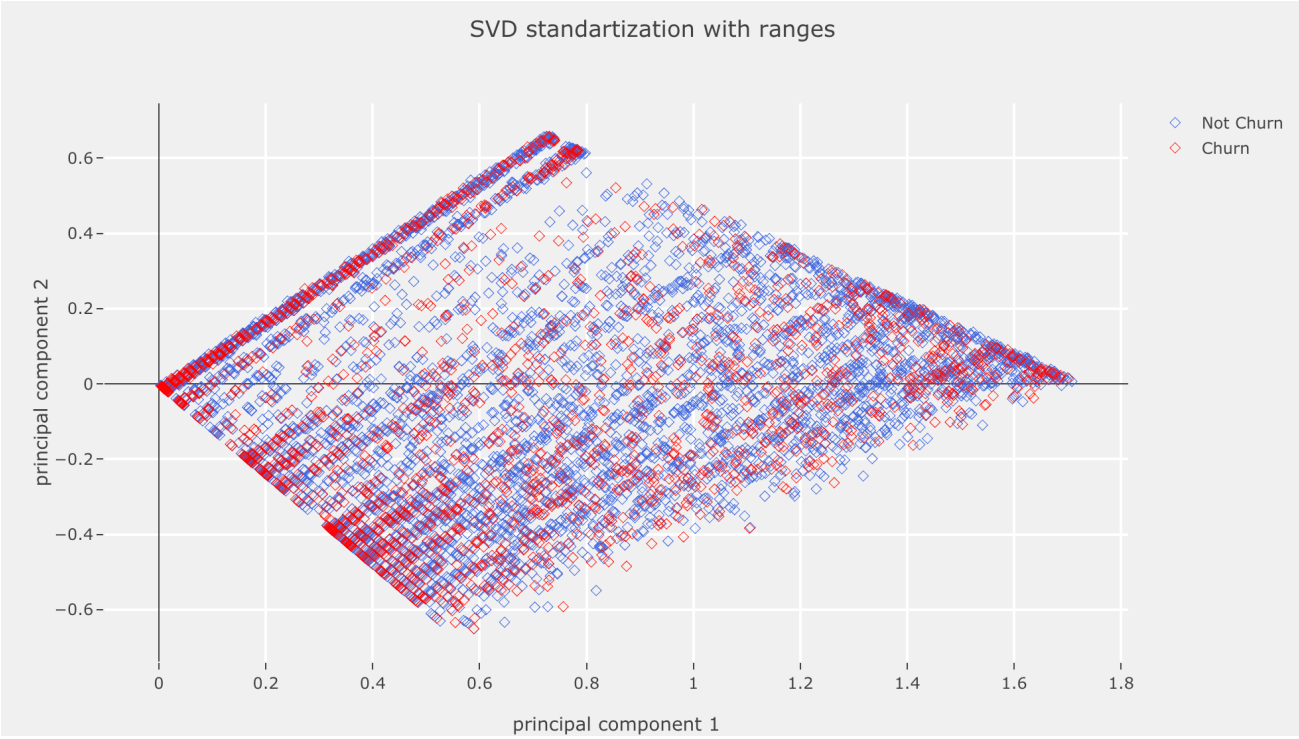
We will use the following features:

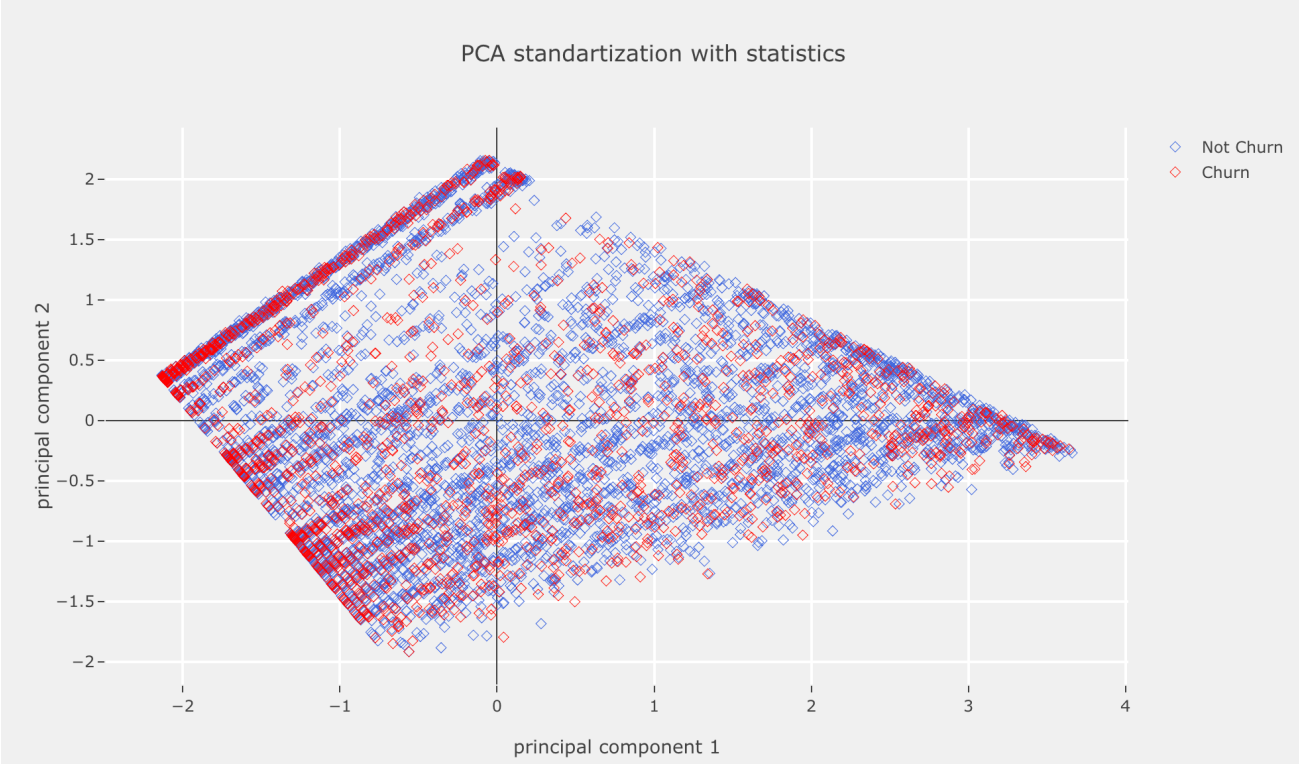
- MonthlyCharges
- TotalCharges
- Tenure

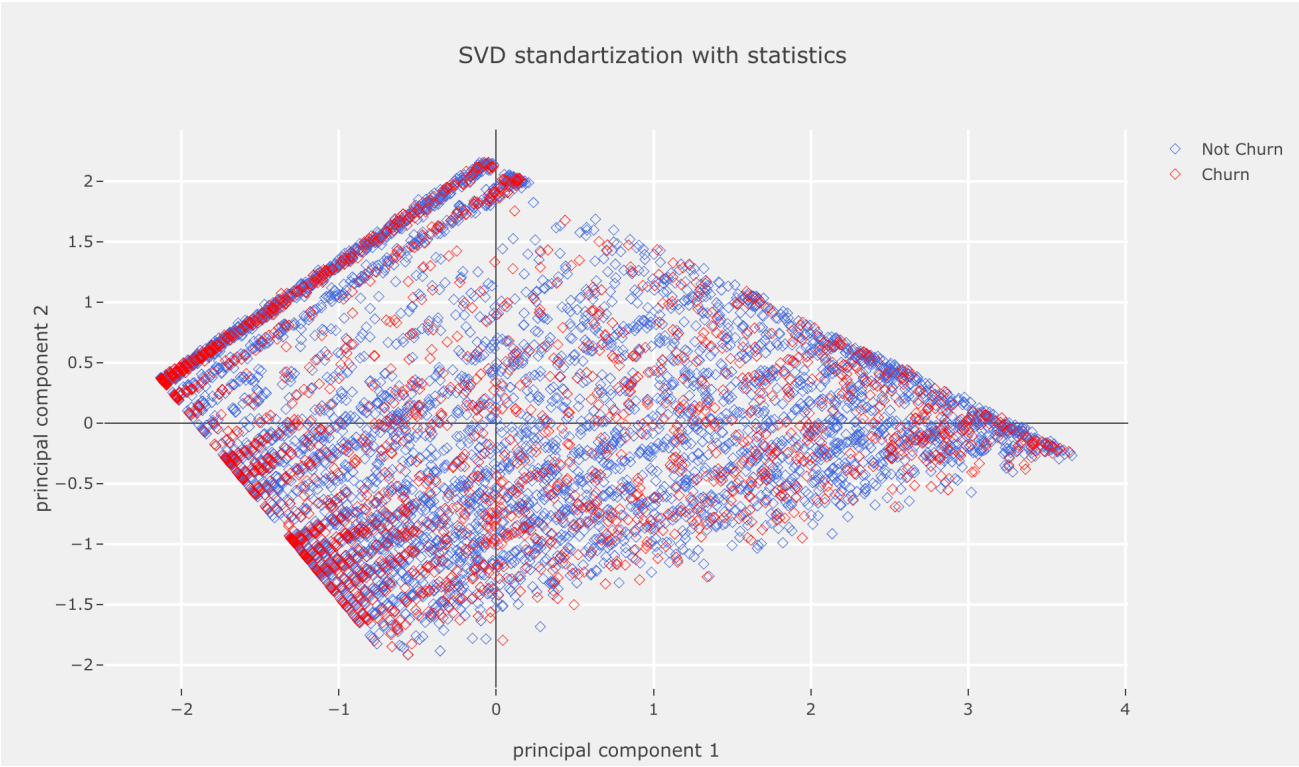
contributions of all the principal components to the data scatter:

naturally: 2.18, 0.76
per cent: 72.6%, 25.35 %


PCA standartization with ranges
Not Churn
Churn
principal component 2
principal component 1
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1


SVD standartization with ranges
Not Churn
Churn
principal component 2
principal component 1
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8



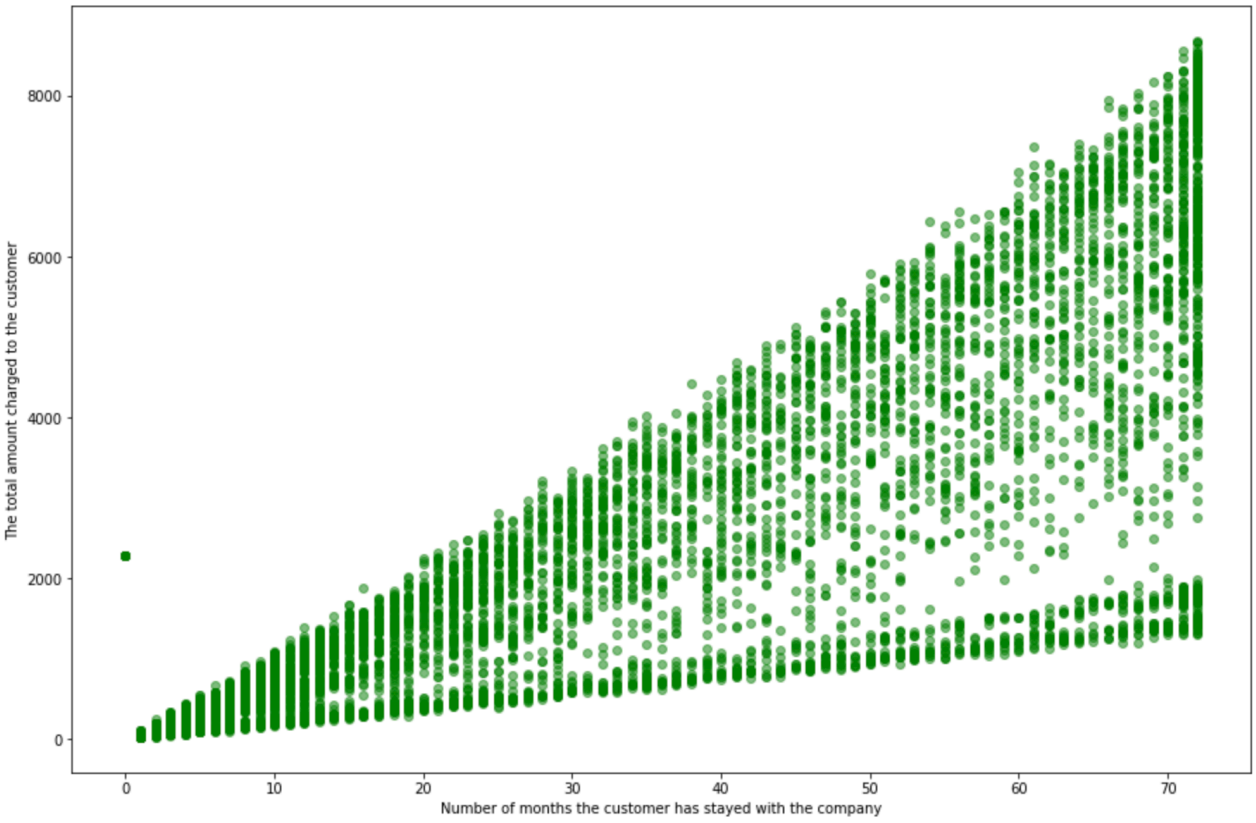


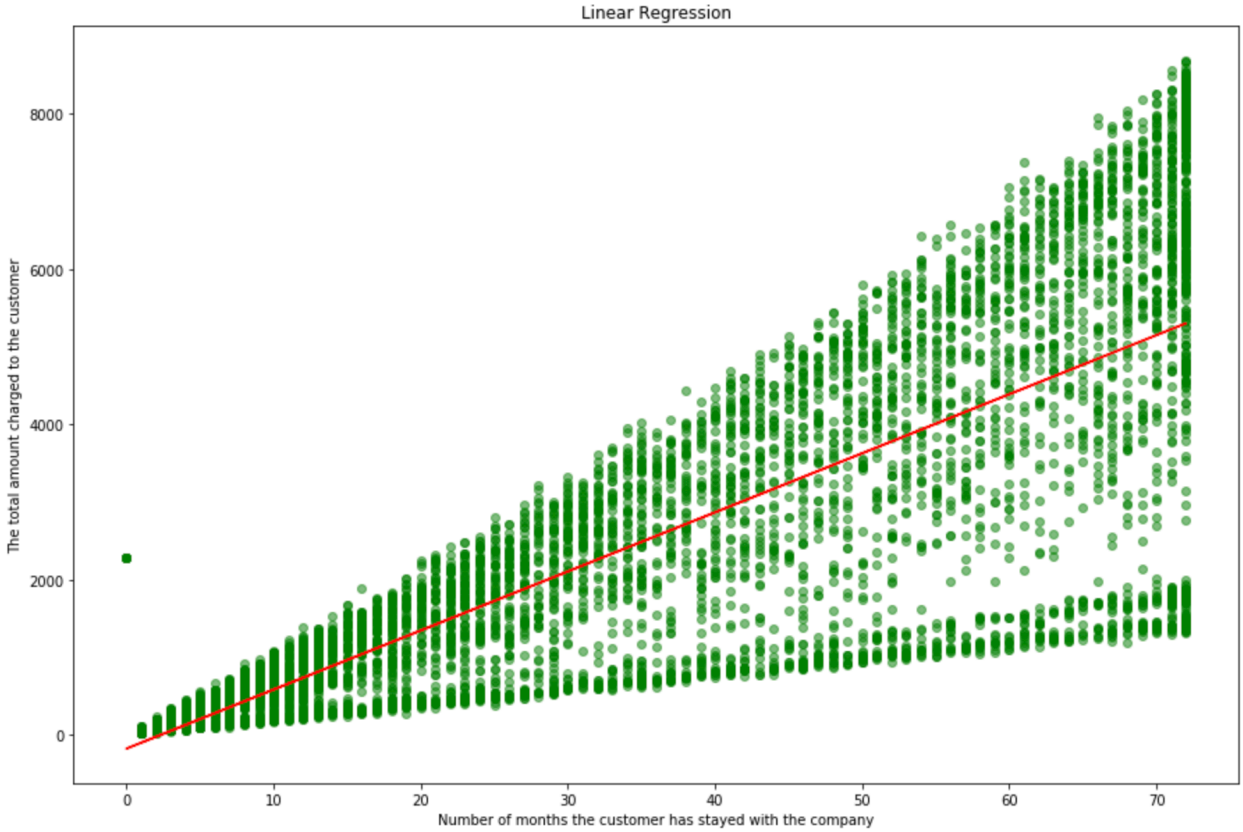
hidden factor behind the selected features: `100`

Task 5

1. Find two features in your dataset with more or less “linear-like” scatterplot.
We will use the following numerical features:

- tenure: Number of months the customer has stayed with the company
- total_charges: The total amount charged to the customer





The slope is positive which means that features are positevely related.

correlation coefficient: 0.82

determinacy coefficient: 0.68

By definition correlation coefficient ranges from 1 to 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of 1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables. In our case it's 0.82 which is much closer to 1 than to 0 so we can look at relation between features as linear-like. The value of correlation coefficient is positive so features are related positively. This seems logical: the longer client stays with the company, the greater is overall amount of money he has spent on company's service.

The coefficient of determination is 0.68 and indicates the proportion of the variance σ(y) taken into account by the linear regression of "tenure" feature over "total_charges" feature. The determinacy coefficient is reasonably large, showing that our linear model explains the dependecy quite well. At the same time, due to quite

large variance of the target variable, linear model is not able to explain the variance of target.

Prediction for the last 3 clients in the dataset:

Id	Target value	Predicted value
7040	657.74	346.45
7041	125.29	306.6
7042	4841.22	6844.5

The prediction is not very accurate.
Mean relative absolute error: 0.97

Mean relative absolute error gives a better score to our model then determinacy coefficient. Obviously, because of high variance of target, determinacy coefficient is quite close to 0.5, while mean relative absolute error is closer to 1.