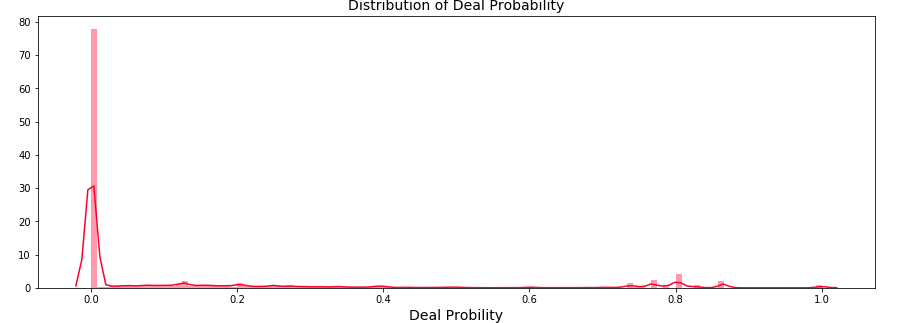
**Data dimentions to deal with:-**

1. **Images**
2. **NLP Russian text data – title and description**
3. **Geogrphy data**
4. **Time series data**

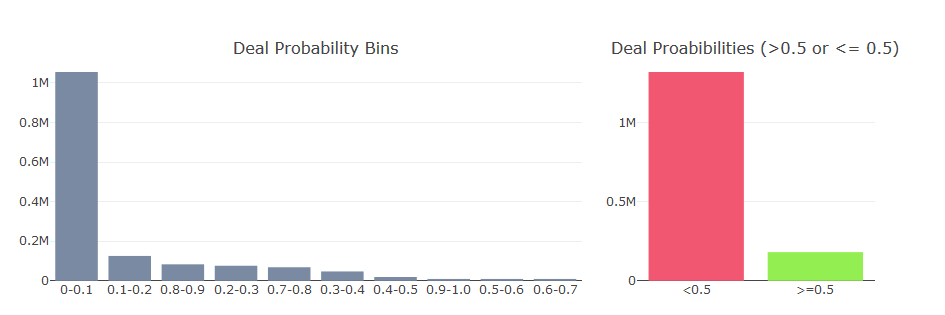


**Inference**

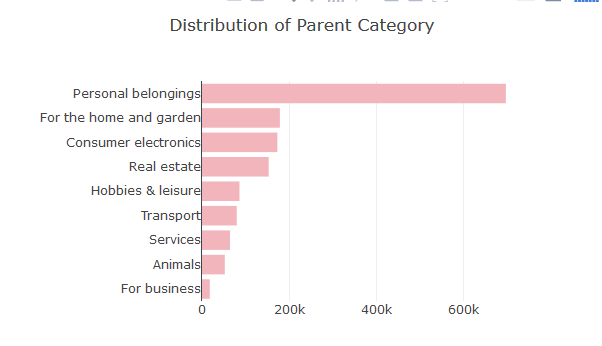
* From the deal probability distribution graph, it is clear that majority of the items have extremely low deal probability, ie. about 78%, while very few values have the deal probability of 0.7 or larger.
* A very small tower is observed near the deal probability of 1.0, indicating that there are some items in the dataset having very high value of deal probability.

**Top 10 Items having deal probability = 1.0 are**

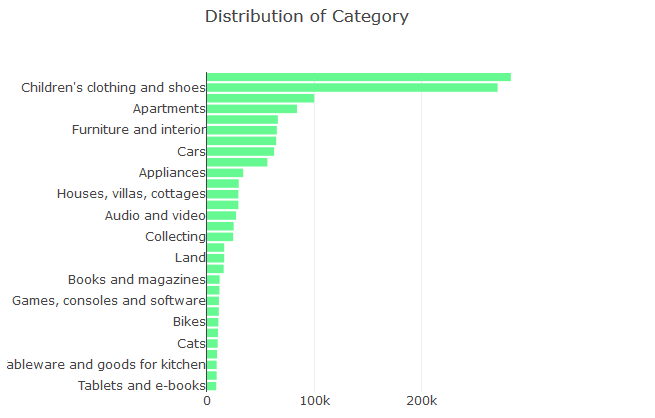
* 92013ca1fe79 | Installation of doors, openings, slopes, arches
* c6239fc67a6f | Nail extension, correction
* 44aa121e4559 | Cargo transportation (long distance), onboard, open
* b16d1b27c975 | Rise of houses
* fe03dbc60ccf | Transportation across the southern Federal district North Caucasian Federal district Crimea



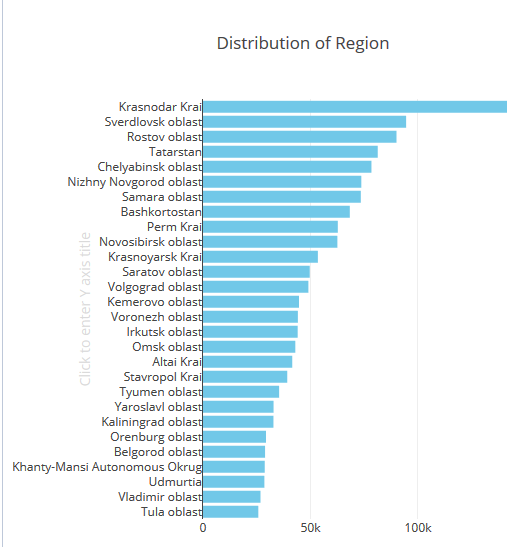
* There are about 1.3 Million items having deal probability less than 0.5, while there are about 182K items having deal probability greater than 0.5
* Most dominant categories are items having deal probability: 0-0.2, 0.8-0.9



* In this dataset, maximum number of items are present under "Personal Belonging" parent category equal to about 700K items.
* Home and Garden, and Consumer Electronic items are also populated in Avito.
* Reletively, fewer number of items which are related to Businesses are present (18K)

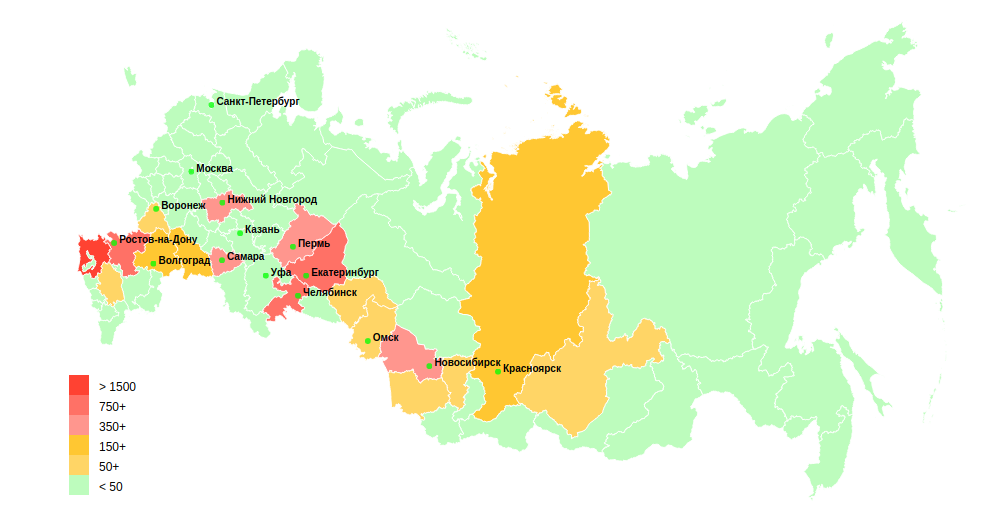


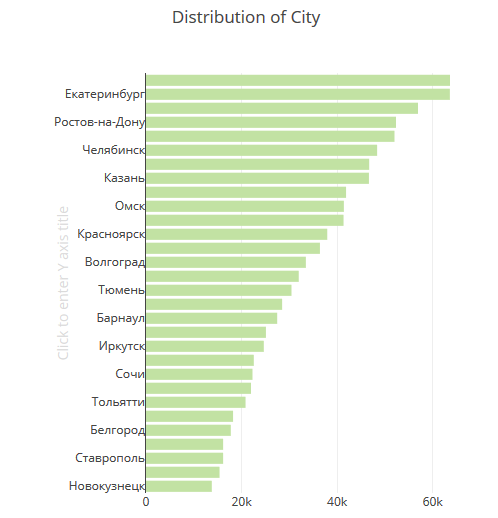
* As also observed in the previous graph that items related to personal belongings are largely present in the dataset. Under these items, Clothing, Accessories, Children Apparels are very common. In total about 550K items are present in these top categories.
* Similar to clothing category, Shoes is another popular category of items present in the dataset, as child shoes and adult shoes items are high in number
* Avito also showcases a large number of advertisments related to child related items such as toys.
* Large number of ads on Avito about Apartments and its essential compontns such as furniture, interor designing can also be found.
* Categories such as plants, Tablets and e-books have relatively lower number of advertisments present on Avito. These might not be performing well on Avito



* The region - Krasnodar Krai tops the table having maximum number of items equal to 141K ie. exteremely high number of ads / items are promoted in this region.
* Sverdlovsk oblast and Rostov Oblast are another top regions having large number of items equal to 94K and 89K respectively

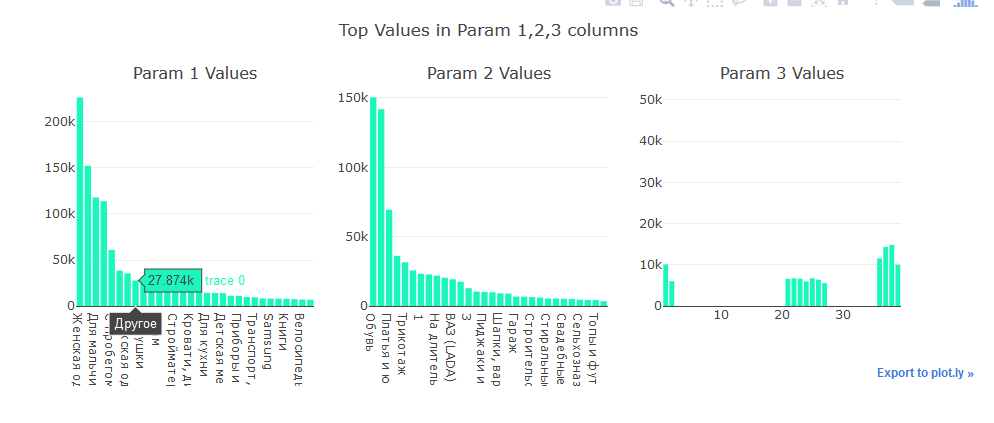
**Visualizing the Regions on the Map**



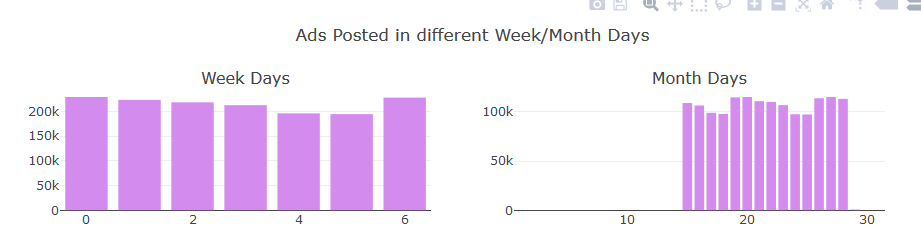


Top cities are: Krasnodar (63638 items), Yekaterinburg, (63602 items), Novosibirsk (56929 items), Rostov-on-don (52323 items)

### Understanding distributions of Param\_1, Param\_2, and Param\_3 in the dataset

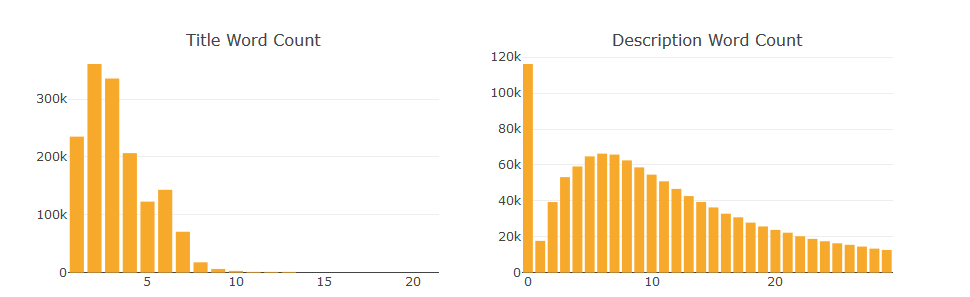


### On which Month Day and Week Day, the items are activated



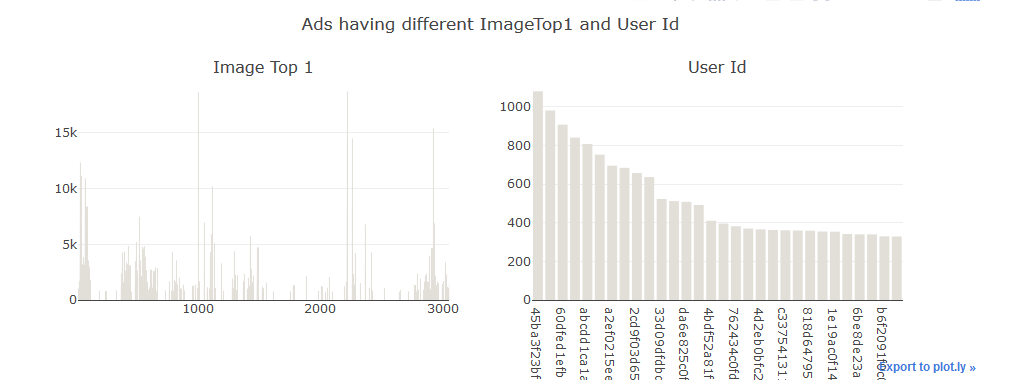
* More number of items are published on Weekends (about 460K in total), while Fridays appear to have lowest number of items
* In the given dataset, maximun number of items were observed on 19th and 20th of the month, while lowest on 17th and 18th. The exact dates from the database are March 19th and 20th of 2017, and lowest - March 17th and March 18th

### What is the Lenght of Item Title and Description



* Items present in the dataset have a range of 1 to 10 words in the title, with title having 2 words in them have the highest number equal to 360K items
* Also, the descriptions of the items takes different values: as low as zero to as high as 25 words. A large number of items with empty descriptions are present in the dataset.
* Items having description containing 6 words are highest while items having description containing 1 word are the lowest equal to 66K and 17K respectively

### Image Top1 and UserId



* User having user\_id = 45ba3f23bf25 have posted maximum number of items (108) followed by user\_id = ee74bccca74f (980 items) and user\_id = 60dfed1efb6e (907 items)
* There are 771769 distincit users who have posted these items while there are 3062 unique image top 1 codes
* Top Image Top 1 codes are 2219 (18739 items), 1002 (18646 items), 2918 (15407 items)

### Distribution of User-Type in the dataset

* Majority of the items are posted by users which belong to Private User type while minority of the items are posted by users which belong to "Shop" user type.
* About 72% of the users are of Private Type while about 5% of the users are of Shop user type

### For what number of Days Ads are Run

### 

* About 65% of the ads/items on Avito are activated for about 2 weeks (ie close to 13 days). There are about 6.9M such ads in the dataset which were activated for 13 days.
* There are many ads in the dataset which were activated for very small period of time ie - for few hours (about 546K ads), probably these were the ads having very poor deal probability
* About 20% of the ads in the datset are are activated not more than a week (ie. less than 7 days)

## Top Brands/Products present in Title

### 

* A large number of items are about Smart Phones and electronic items
* Some noticable Smartphones: **Iphone 5s, Iphone 6s, Samsung Galaxy, Sony Xperia, Xiaomi, Redmi, Nokia**
* Some noticable Electronics Terms Used: **Xbox, PS3, Lenovo, Asus**
* Some noticable Clothing Brands: **Addidas, Reebok, Zara**

## Multi-Variate Analysis

### Correlation in the dataset

### 

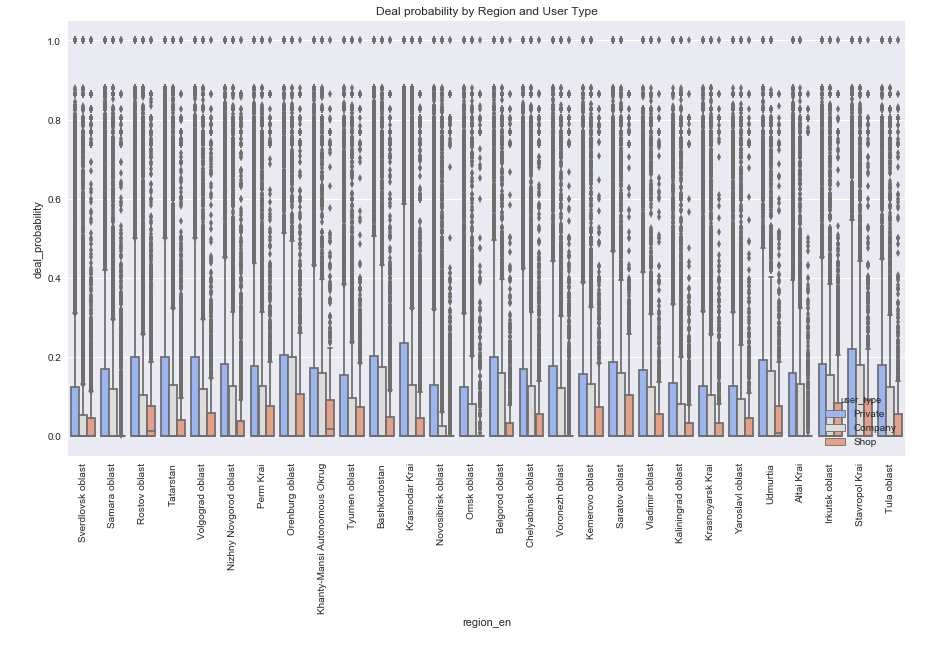
* From the correlation graph, Deal probability shows good correlation with description len and title len. These may by important features while modelling.
* Item Seq Number is also correlated mildly with title len, description len, and Image Top 1
* Image Top 1 code shows a good correlation with deal probability and length of title and description

### Deal Probability by Parent Category and User Type

### 

* From the above boxplot graph, Users having "shop" user types are only related with items of Real Estate, Transport, and Services categories. Service Category is the one having largest deal probability, while Real estate have lowest.
* Users having "Company" as user type also have highest deal probability in the Services category followed by Transport and Animals. Unlike Shop user types, Company user types are related with almost all parent categoies
* Users having "Private" as user type are scattered in the range of 0.1 to 0.6 in the Services category.

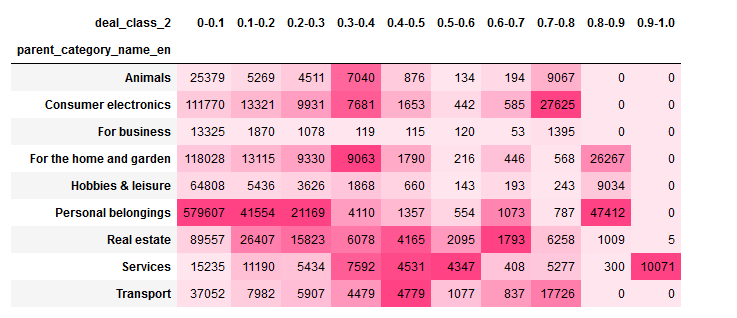
### Deal Probability by Region and User Type



* Above box plot shows that Items which are posted in Orenburg Oblast and Saratov Oblast and which belongs to Shop user type has higest range of deal probabilities.
* For the items from Company User type users, maximum variation is shown in the Stavropol Krai and Orenburg Oblast.

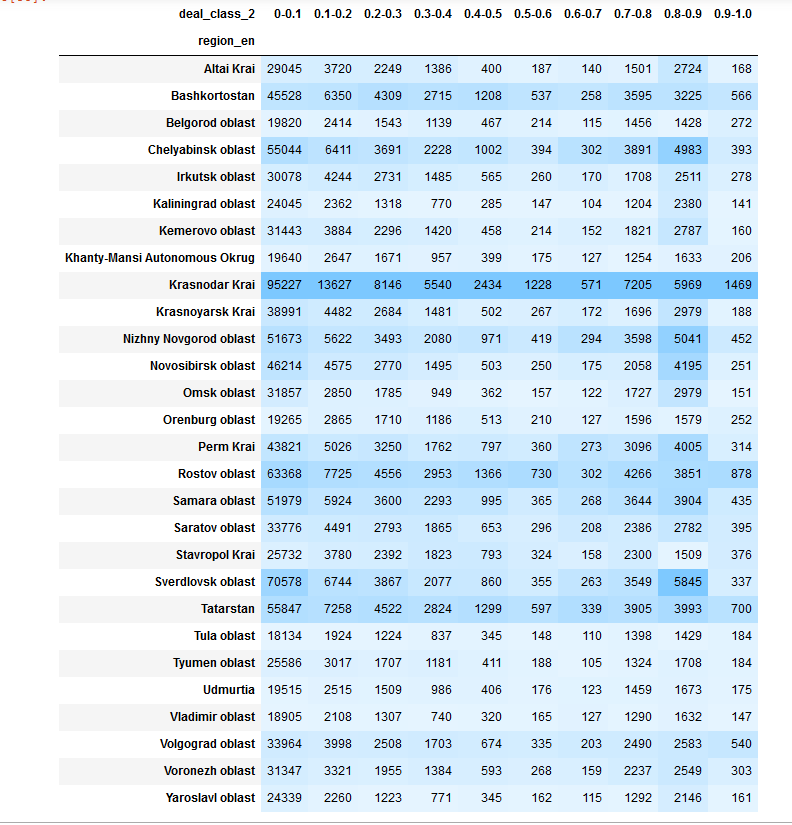
Lets Understand the Mean price of Items by Different Variable Groups

### How many items are there under different Parent Category and different Deal Class

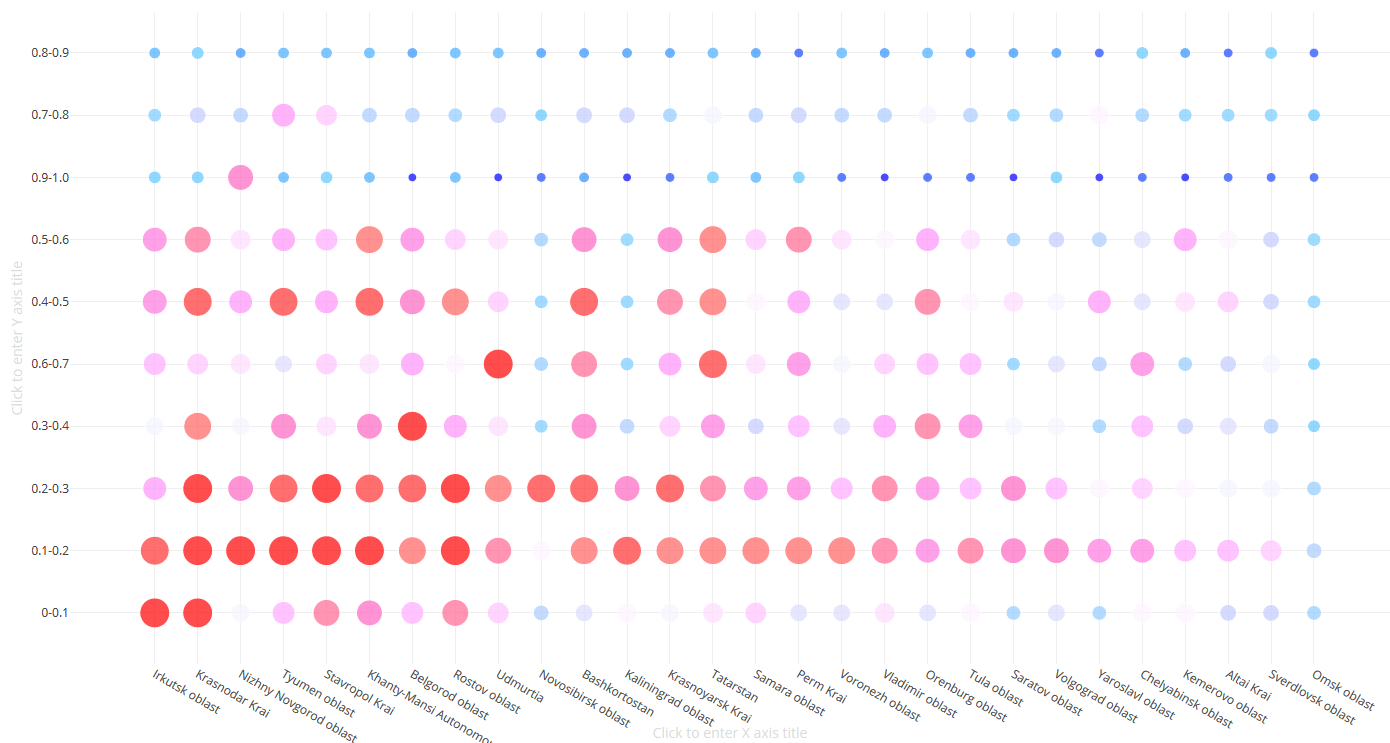


* A large number of Consumer Electronics items have high deal probabiliy (0.7 - 0.8), Personal Belongings category which dominates the dataset also contains higher number of items having deal probability close to 1.0
* Home and Garden related category also contains items having high deal probability.
* Only services category has the items having deal probability greater than 0.9

### Items from different Regions and Different Deal Probability

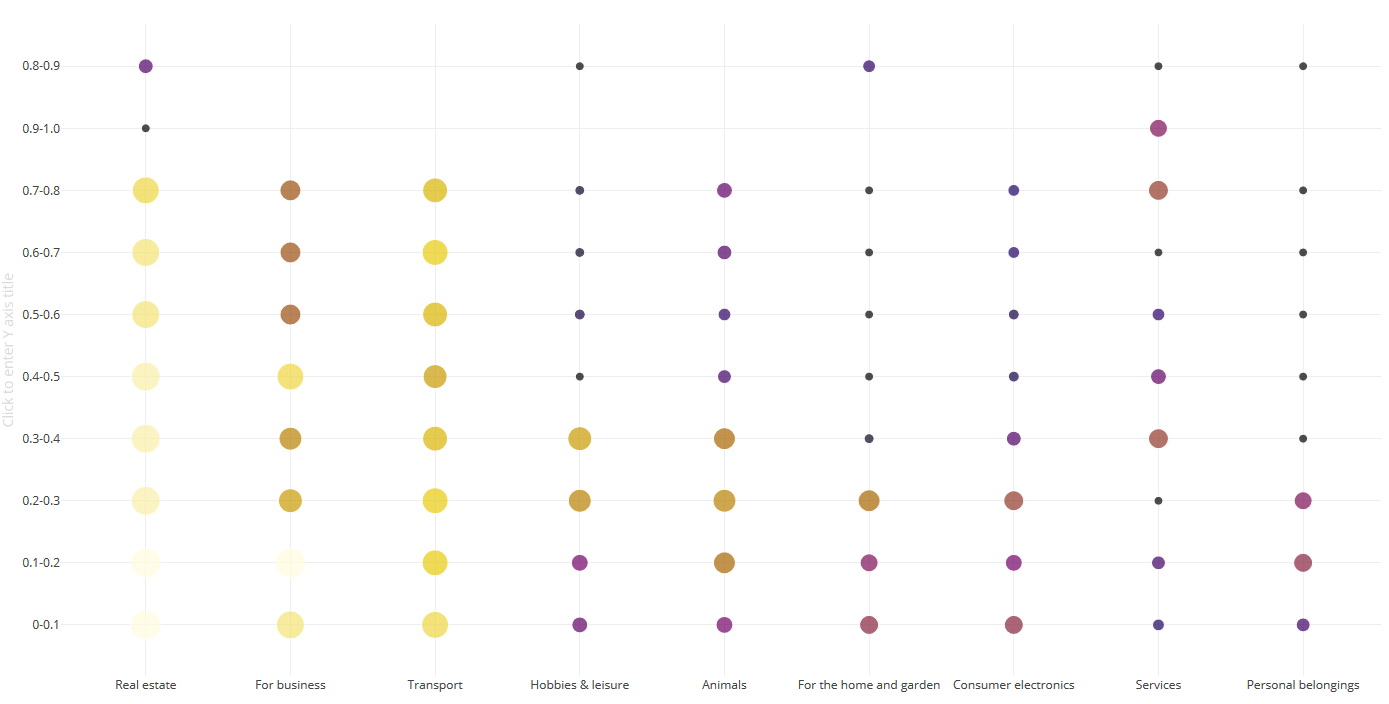


### Different Regions and the mean price of items having different deal probability



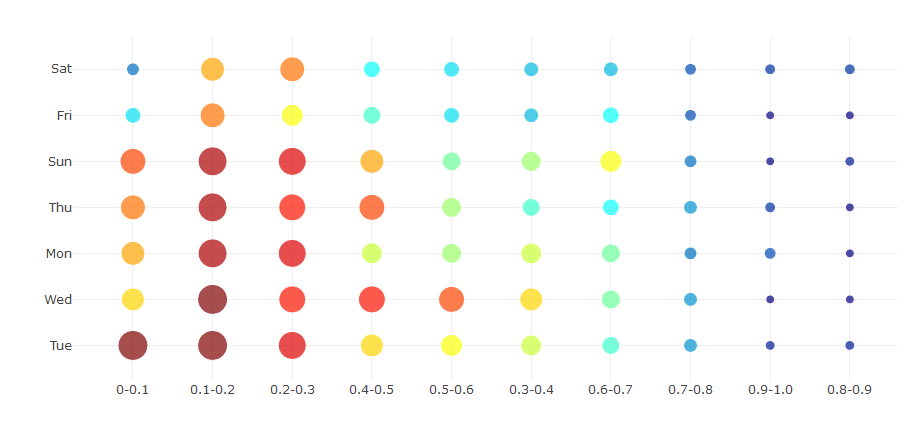
* Items from Kransnodar Krai Region, Irkutsk Oblast, and Nizhny Novgorod Oblast having low deal probability (less than 0.2) have maximum average price than any other region.
* Items with higher mean price and having higher deal percentage (0.7 - 0.8) belongs to Tyumen Oblast and Stavropoi Krai regions.
* Kransnodar Krai (having highest number of items) has different varities of deal percentages of items with higher mean prices

### What is the Mean Price of Items having different Deal Probability and Parent Category



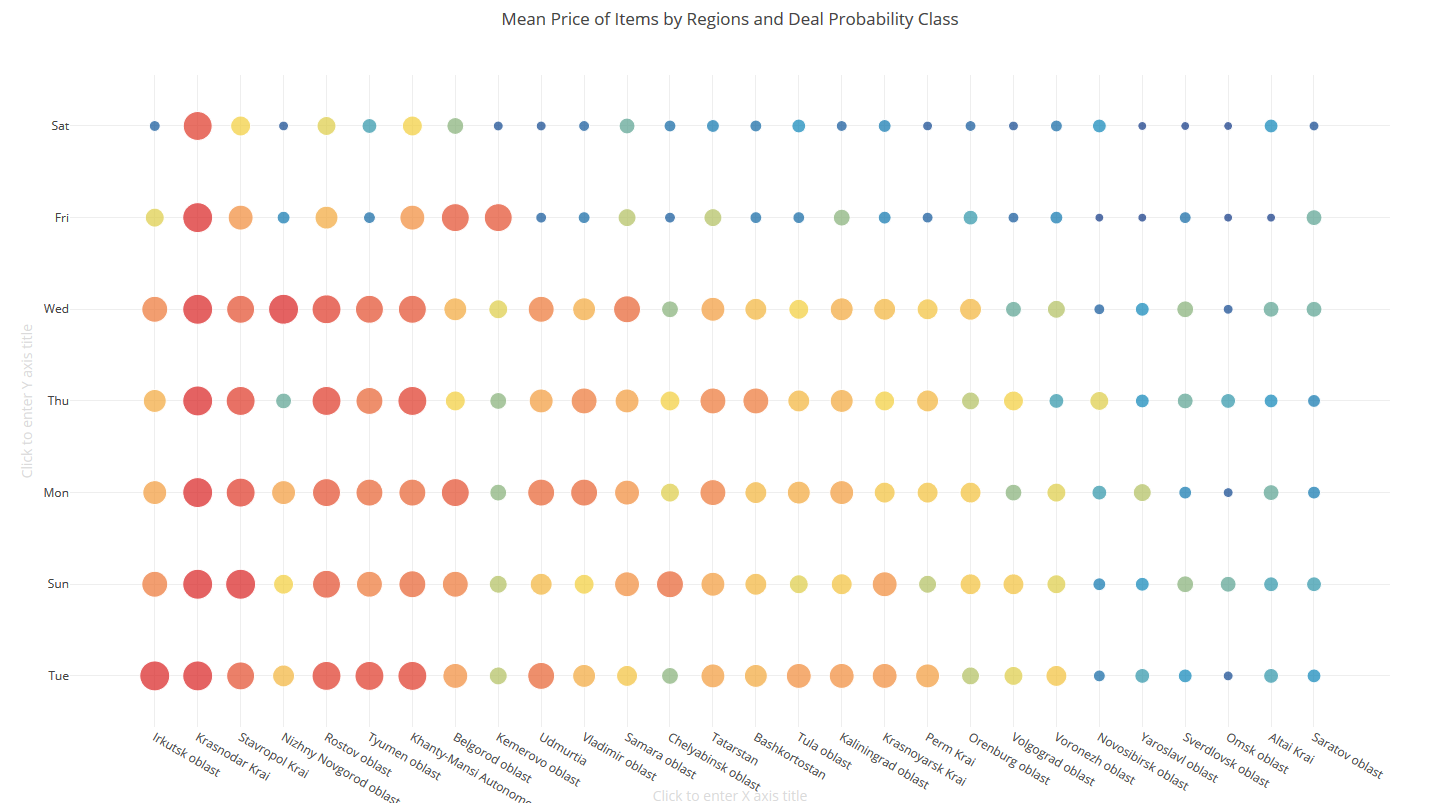
* Real Estate and Transport category generally contains items having higher average prices
* Personal beloningns category do not contains items having higher average prices with any value of the deal probability
* Items having higher deal probability and higher price belongs to categories - home and garden, and real estate

### Deal Class values on different Week Days and their Mean Price



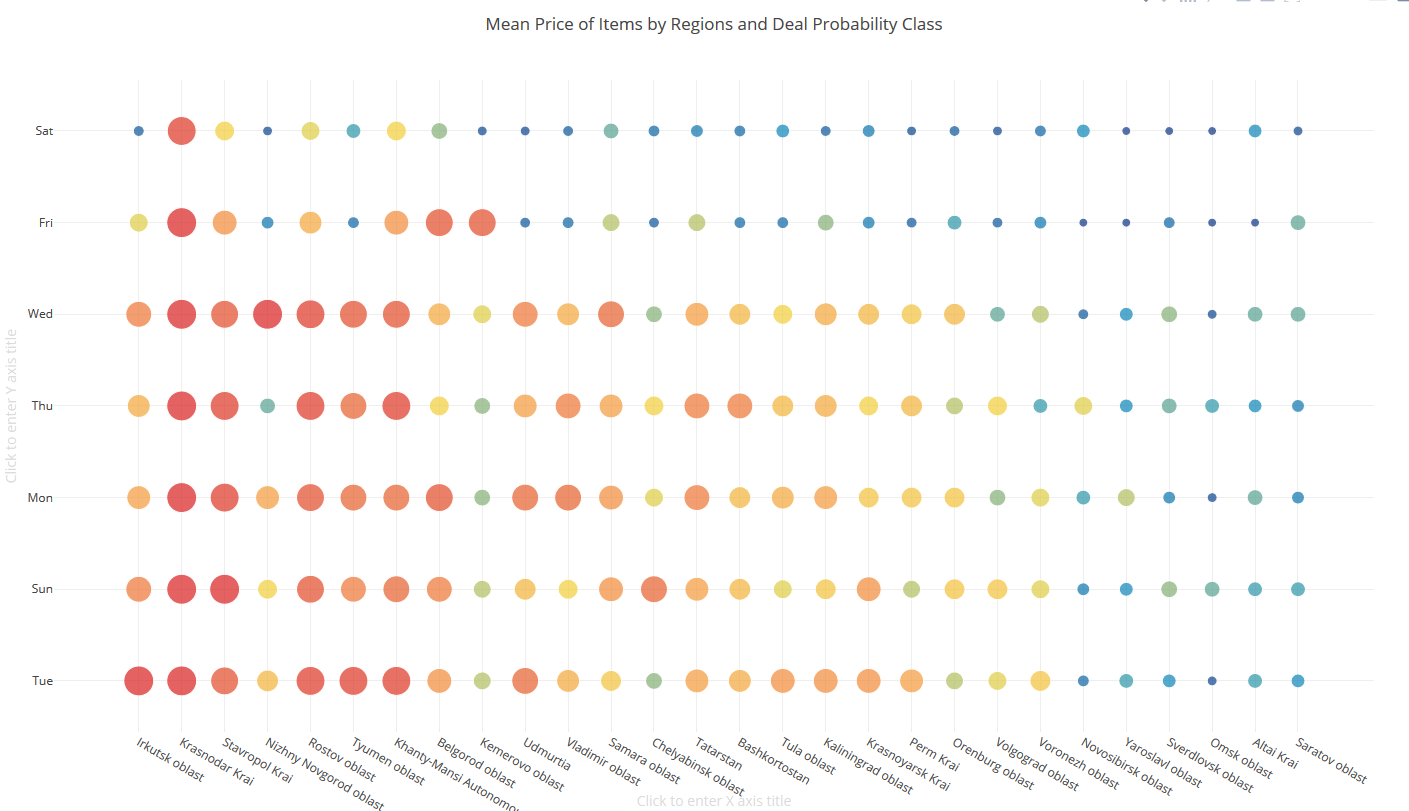
* Average price of Items with high deal percentage (0.9+) and which were posted on Saturdays and Mondays has the higher mean price than any other day
* Maximum price of items having low deal percentage (<0.1) were posted on tuesdays

### Different Regions, Different Day of the Week and Value of Mean Price of Items

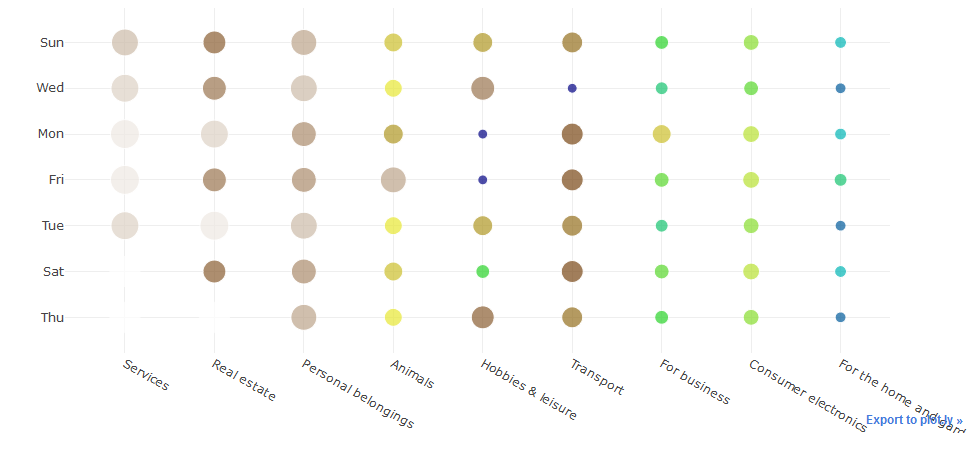


Tuesdays and Wednesdays have most items having higer prices in different regions than any other day while saturdays have lowest

### Week Day and Region vise Value of Max Price of Items

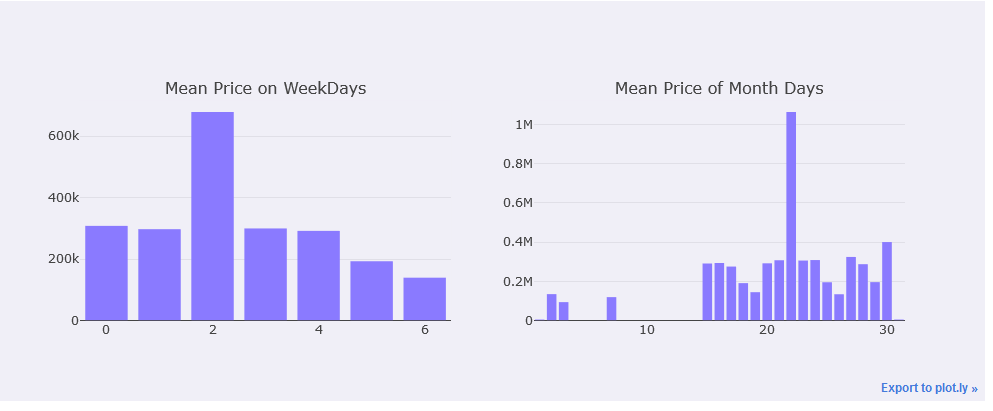


### Max value of Deal Proabability by different Regions and Day of the Week



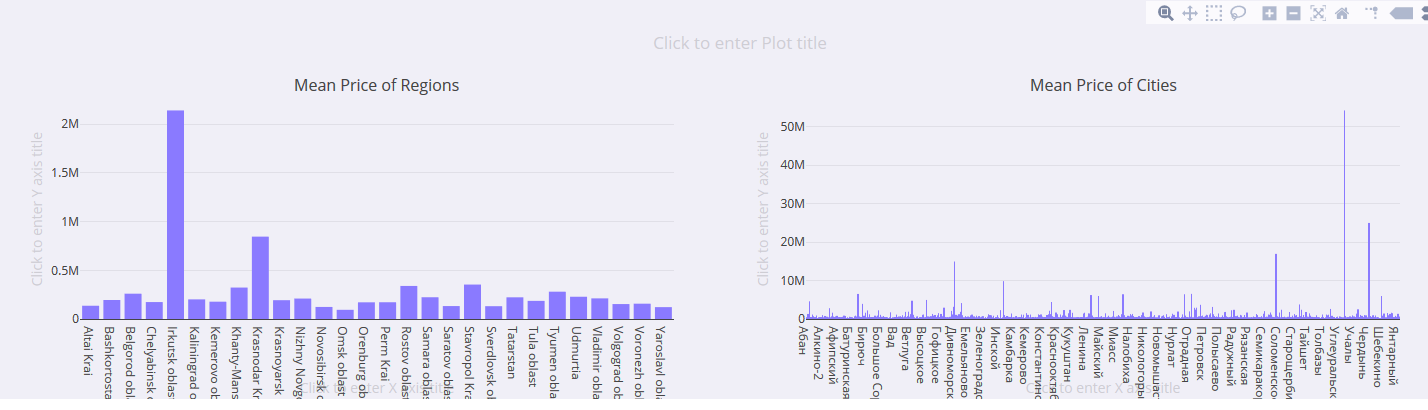
## What are the charateristics of Items having High or Low Mean Price

### Mean Price of Items by Week Day and Month Day



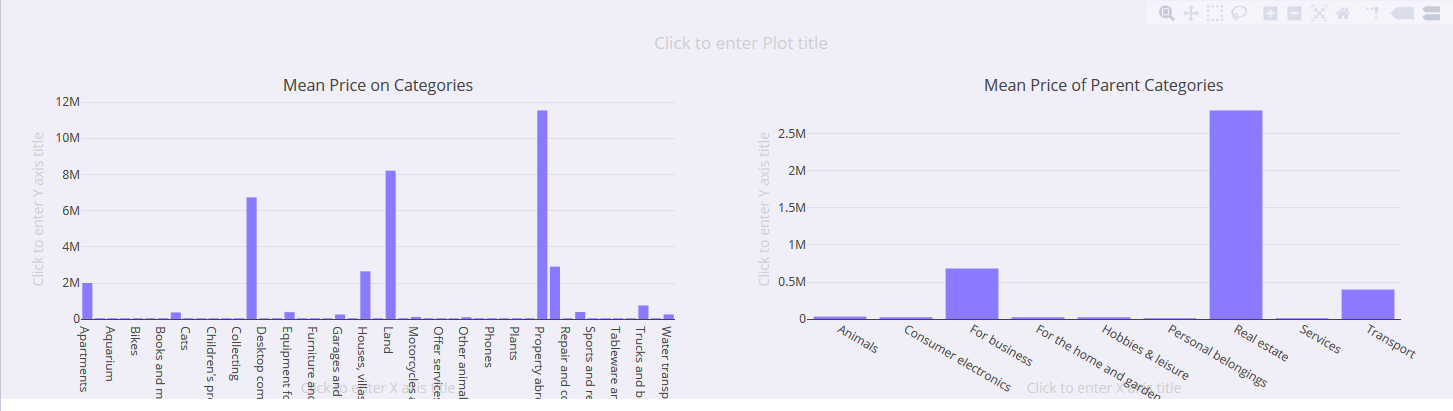
Items with maximum price were promoted on 22nd date of the month ie. 22 March 2017, while the cheapest items was promoted on 3rd of March 2017

### Mean price of items by Regions and Cities



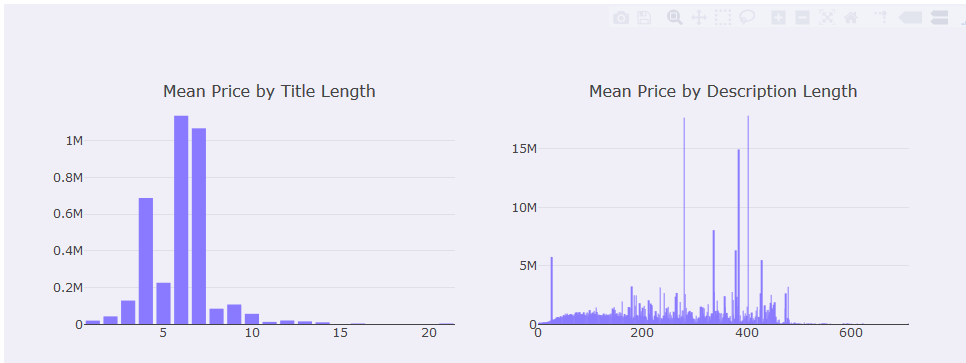
Items having maximum mean price (2.14M) were promoted in Irutsk Oblast Region and Krasnoyarask Krai

### Mean Price of Items by Categories and Parent Categories



* As expected, Real Estate items have highest mean items followed by Business and Transport
* Property Abroa, Land, and Commerical Property have highest mean price

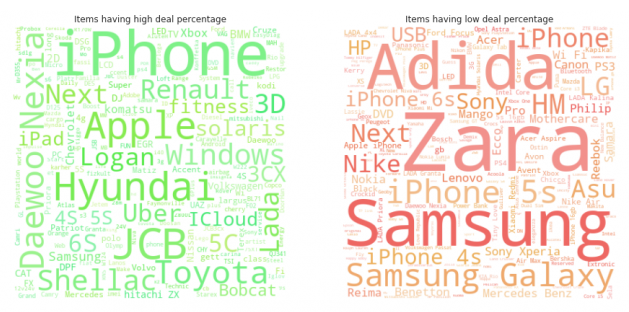
### Mean Price by Count of Words present in Title and Description



* Longer titles have higher price, titles having 6-7 words have very high mean price
* Descriptions with about 400 words have high mean price of about 15M

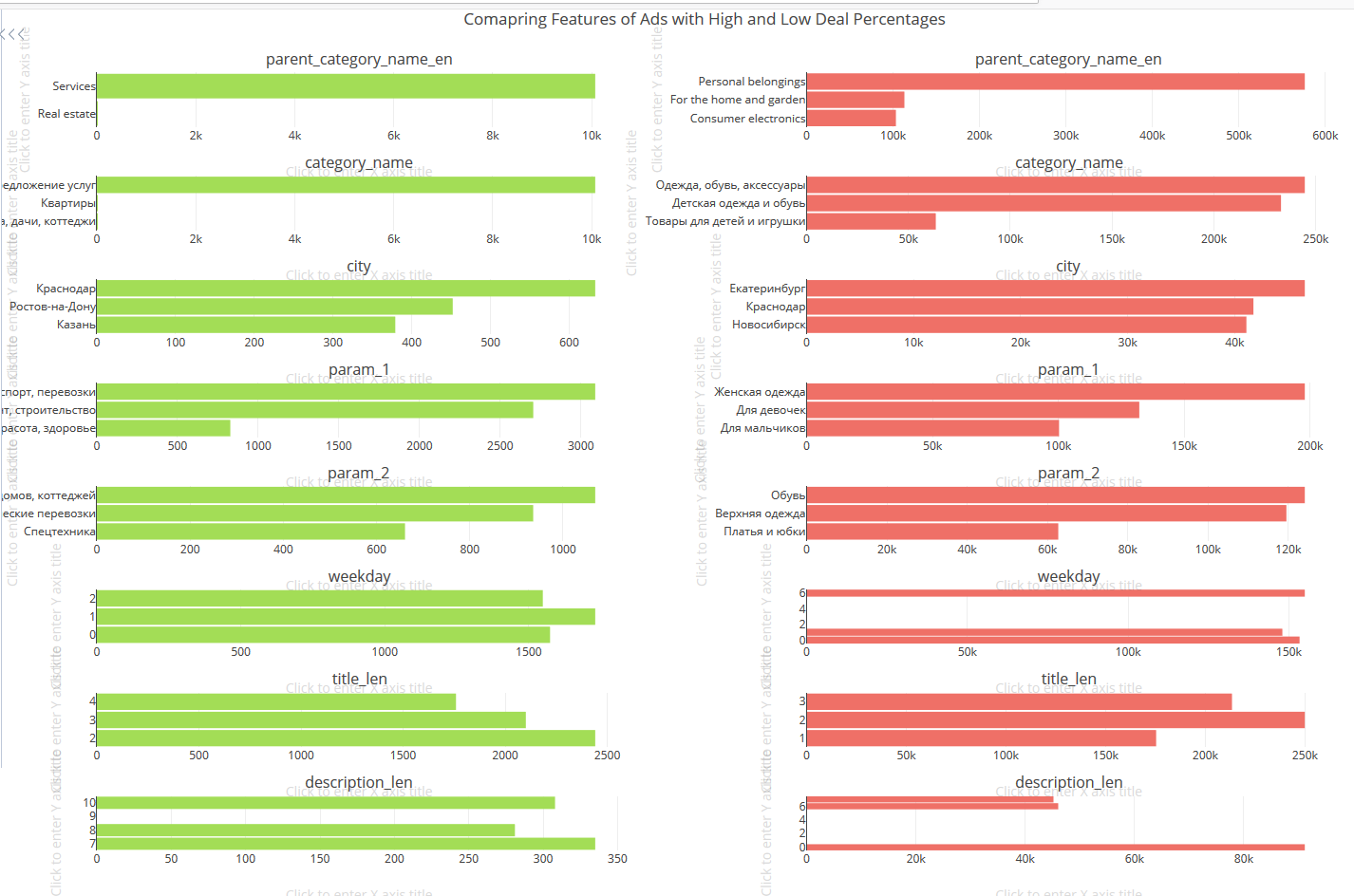
## What are the characteristics of Items having Very High (or Very Low) Deal Percentage

### Words present in the title



* Interesting to see that there are items such as Iphone, Hyuanday, JCB, Toyota which have high deal percentage and they are not present in titles of items having low deal percentage
* Brands such as Samsung, Zara, and Adiddas are from the item sets having low deal probabilities.

### Comparing all other features of Items with Low and High deal percentages



### Top N-grams used in descriptions

### One of the charateristics of good performing ads (having high deal probability) is that they clearly mention about about item such as Roofing Work , Apartment, Quality, Construction Works. While poor performing ads talk about the additional details such as - "excellent condition", "working condition", "bargain", "walking distance etc"

### 