Разведовательный анализ данных

Загрузка данных

Выведим описательные статистики для данных из СберИндекс

summary(df$Индекс.БП)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 9.40 49.20 53.40 51.71 57.10 71.10

summary(df$Индекс.ПА)

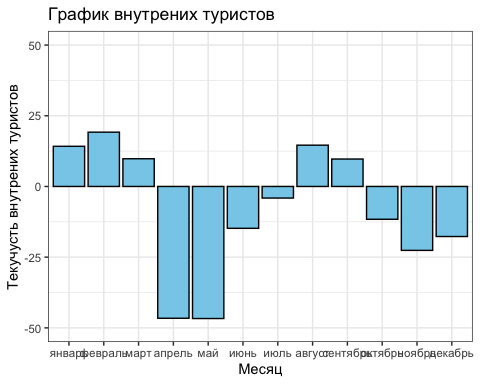
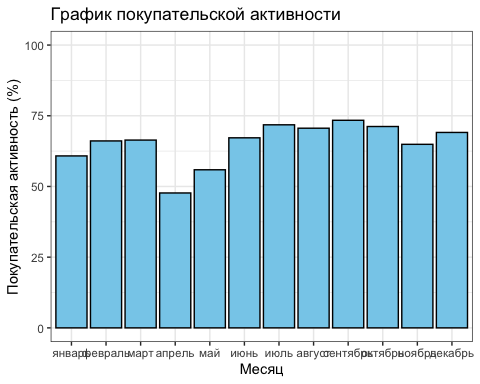
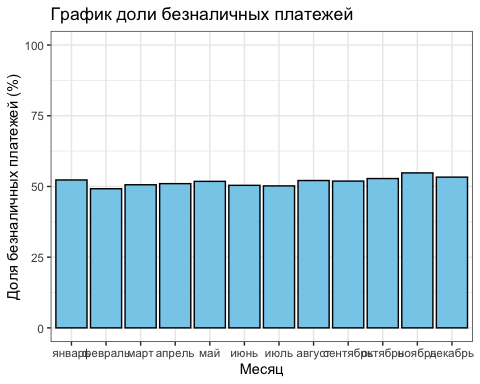
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 34.00 61.00 67.00 65.43 71.00 88.00

summary(df$Количество.ВТ)

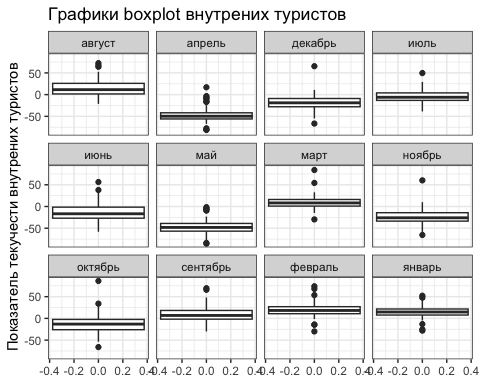
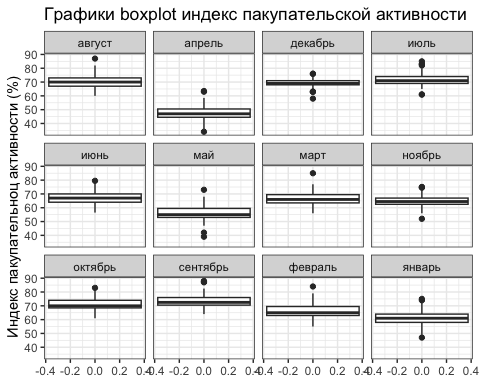
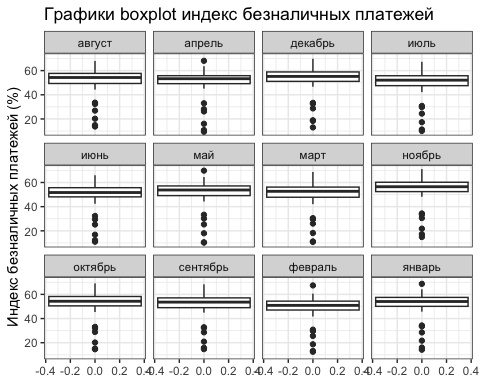
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -85.470 -28.052 -5.125 -8.061 11.162 86.140

Согласно описптельным статистикам видно, что: 1. индекс безналичных платежей в среднем по строне равен, чуть больше 50% (52%), но в некоторых регеонах он возрастает до 70% и может упасть до 10%. Это значит, что в некоторых регионах люди очень мало используют безналичный способ оплаты (возможно такой способ оплаты недоступен), а в некоторых очень активно, но в среднем по стране чуть больше половины людей производит оплату безналичным способом. 2. Индекс потребительской активности всреднем равен 65%, значит что потребители в среднем покупают товаров на 65% от самого активного дня в году (30 декабря), что думаю неплохо. Также есть месеца когда индекс потребительской активности возрастает до 88% и падает до 34%. 3. Что касается количества внутренних туристов, то среднее количество равно примерно минус 8%, знак минус говорит о том что в среднм количество туристов уменьшилось по сравнению с предыдущим годом.

Построим графики для показателей СберИндекс в среднем по России, по месяцам. Для этого создадим дополнительный датафрейм с данными по месяцам в среднем по России.

 1. График доли безналичных платежей в среднем по России показывает, что люди каждый месяц оплачивают безналичным способом чуть больше половины всех покупок. 2. График покупательской способности показывает, что у людей покупательская активность падает в апреле и мая. Наилучшие месяцы по пукапательской активности это начиная с июля и почти до конца года. 3. Что касается графика движения внутрених туристов, то очень сильно просел этот показатель в апреле и мая по сравнению с предыдущим годом, также спад наблюдается и на конец года. Но начало года показало увеличения количества внутрених туристов. Так же конец лето и начало осень идет увеличения притока внутрених туристов.

Построим графики boxplot для показателей СберИндекс.



Выведим нетипичные значения в отдельные датафреймы, соответственно по показателям СберИндекса.

После анализа датафрейма с нетепичными значениями можно предположить сдледующее: 1. В датафрейме с индексом безналичных платежей, видим, что к ним относяться большенство регионов кавказа (причем в выборку попали регионы полностью по всем месяцам) и показатель всегдла ниже перевого квантиля. На мой взгляд не стоит относить эти значения к выбросам и удалять их, по следующим причинам: либо в этих регионах недостаточно развита система безналичных платежей, либо это связано с менталитетом местного населения, предпочитающих расплачиваться живыми деньгами. 2. В датафрейме с индексом покупательской активности, значения которые меньше первого квантиля приходяться на апрель и май. Если мы посмотрим на график покупательской активности, то увидим, что в эти месяца идет спад этого показателя. Можно предположить, что значения неявляются выбрасами удалять их также нестоит, а обусловленно спадом покупательной активности. То же самое и со значениями превышающие третий квантиль, выподают на сентябрь месяц, что подтверждает график покупательской активности. 3. Значения в датафрейме количества внутрених туристов, тоже нестоит относить к выбросам, так как занчения больше третьего квантиля и они приходяться на те месяца когда идет рост движения внутрених туристов, что подтверждает график количества внутрених туристов. Если говорить в целов нетипичные значения отличаются примерно на 1-5%, на мой взгляд нестоит удалять их. Они неповлияют на модель, а из-за их удаления можем потерять много полезной информации.

Выберем топ 30 регионов ледирующих по индексу покупательской спосбности.

Выберем топ 30 регионов лидирующих по индексу безналичных платежей.

Выберем те регионы которые лидтруют по индексу покупательской способности и индексу безналичных платежей.