Разведывательный анализ данных

## Выбор регионов, лидирующих по потребительской активности, безналичным платежам и внутреннему туризму

Загрузка данных

Выведем 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)

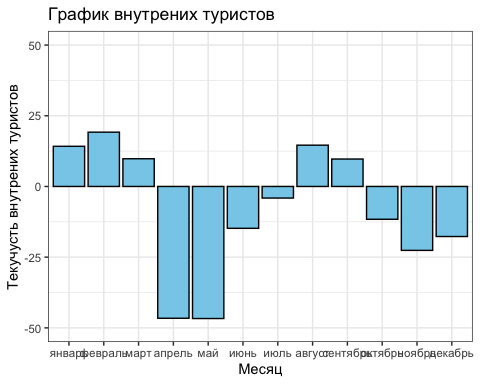
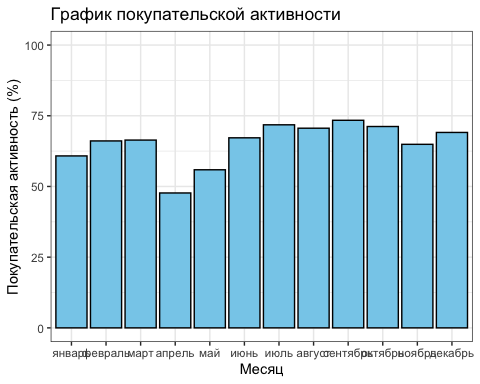
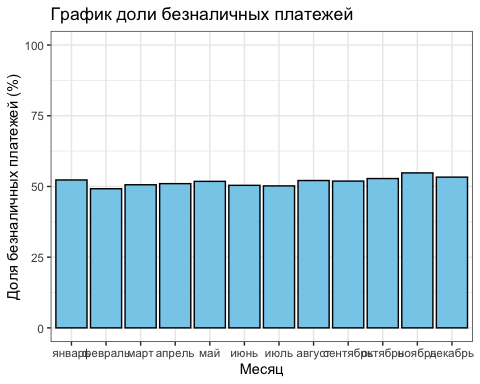
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 34.00 61.00 67.00 65.43 71.00 88.00

summary(df$Количество.ВТ)

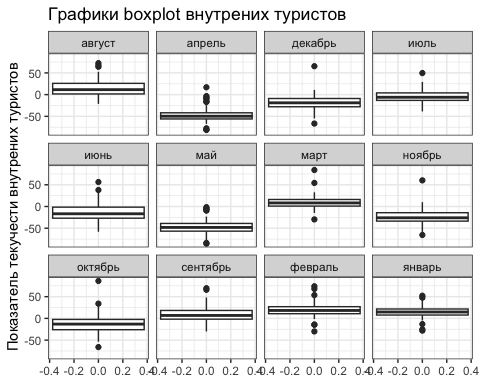
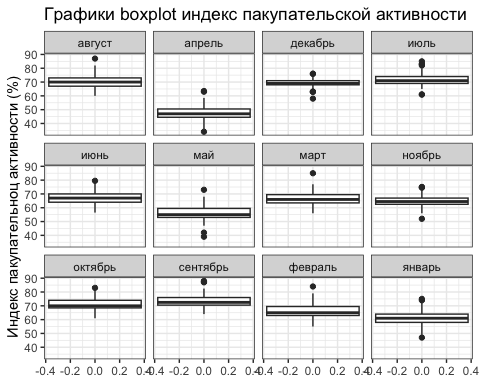
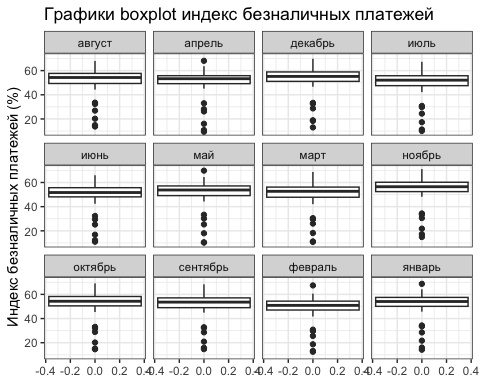
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -85.470 -28.052 -5.125 -8.061 11.162 86.140

Согласно описательным статистикам видно, что: - индекс безналичных платежей в среднем по стране равен, чуть больше 50% (52%), но в некоторых регионах 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. - Индекс потребительской активности в среднем равен 65%, значит что потребители в среднем покупают товаров на 65% от самого активного дня в году (30 декабря), что думаю неплохо. Также есть месяца когда индекс потребительской активности возрастает до 88% и падает до 34%. - Что касается количества внутренних туристов, то среднее количество равно примерно минус 8%, знак минус говорит о том что в среднем 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.

 - График доли безналичных платежей в среднем по России показывает, что люди каждый месяц оплачивают безналичным способом чуть больше половины всех покупок. - 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. - Что касается графика движения внутренних 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 внутренних туристов. Так же конец лето и начало осень идет увеличения притока внутренних туристов.

3 .Построим графики boxplot для показателей СберИндекс.



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

## # A tibble: 36 × 16  
## Регион Год Месяц Квартал `Число абонентов` `Среднемесячная з.п.`  
## <chr> <dbl> <chr> <chr> <dbl> <dbl>  
## 1 астраханская обл… 2020 апре… кварта… 745 35800.  
## 2 белгородская обл… 2020 сент… кварта… 15804 36501.  
## 3 вологодская обла… 2020 апре… кварта… 2230 40288.  
## 4 воронежская обла… 2020 сент… кварта… 5901 35744.  
## 5 ивановская облас… 2020 апре… кварта… 2518 26770.  
## 6 кабардино-балкар… 2020 апре… кварта… 60 28006.  
## 7 калининградская … 2020 авгу… кварта… 828 36102   
## 8 калининградская … 2020 сент… кварта… 828 36258.  
## 9 костромская обла… 2020 апре… кварта… 30 28995   
## 10 краснодарский кр… 2020 апре… кварта… 10206 36885.  
## # ℹ 26 more rows  
## # ℹ 10 more variables: `Уровень безработицы` <dbl>, Индекс.БП <dbl>,  
## # Индекс.ПА <dbl>, Количество.ВТ <dbl>, `Всего одобренных заявок` <chr>,  
## # `Онлайн-заявки` <dbl>, `Офлайн-заявки` <dbl>,  
## # `Всего ипотечных сделок` <chr>, `Доля сделок, первичка` <dbl>,  
## # `Доля сделок, вторичка` <dbl>

## # A tibble: 76 × 16  
## Регион Год Месяц Квартал `Число абонентов` `Среднемесячная з.п.`  
## <chr> <dbl> <chr> <chr> <dbl> <dbl>  
## 1 кабардино-балкар… 2020 янва… кварта… 57 27758.  
## 2 кабардино-балкар… 2020 февр… кварта… 57 27761.  
## 3 кабардино-балкар… 2020 март кварта… 57 27045.  
## 4 кабардино-балкар… 2020 апре… кварта… 60 28006.  
## 5 кабардино-балкар… 2020 май кварта… 60 31536.  
## 6 кабардино-балкар… 2020 июнь кварта… 60 35228.  
## 7 кабардино-балкар… 2020 июль кварта… 65 27222.  
## 8 кабардино-балкар… 2020 авгу… кварта… 65 27089.  
## 9 кабардино-балкар… 2020 сент… кварта… 65 28665.  
## 10 кабардино-балкар… 2020 октя… кварта… 71 29742.  
## # ℹ 66 more rows  
## # ℹ 10 more variables: `Уровень безработицы` <dbl>, Индекс.БП <dbl>,  
## # Индекс.ПА <dbl>, Количество.ВТ <dbl>, `Всего одобренных заявок` <chr>,  
## # `Онлайн-заявки` <dbl>, `Офлайн-заявки` <dbl>,  
## # `Всего ипотечных сделок` <chr>, `Доля сделок, первичка` <dbl>,  
## # `Доля сделок, вторичка` <dbl>

## # A tibble: 4 × 16  
## Регион Год Месяц Квартал `Число абонентов` `Среднемесячная з.п.`  
## <chr> <dbl> <chr> <chr> <dbl> <dbl>  
## 1 владимирская обла… 2020 февр… кварта… 1437 32168.  
## 2 владимирская обла… 2020 март кварта… 1437 33768.  
## 3 владимирская обла… 2020 авгу… кварта… 2275 32390.  
## 4 владимирская обла… 2020 октя… кварта… 2296 34611.  
## # ℹ 10 more variables: `Уровень безработицы` <dbl>, Индекс.БП <dbl>,  
## # Индекс.ПА <dbl>, Количество.ВТ <dbl>, `Всего одобренных заявок` <chr>,  
## # `Онлайн-заявки` <dbl>, `Офлайн-заявки` <dbl>,  
## # `Всего ипотечных сделок` <chr>, `Доля сделок, первичка` <dbl>,  
## # `Доля сделок, вторичка` <dbl>

После анализа датафрейма с нетепичными значениями можно предположить сдледующее: - 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. - 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. - 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 лидируют по индексу покупательской способности и индексу безналичных платежей.

Выберем строки из датафрейма df которые соответствуют выборки лидирующих регионов

9,10. Отфильтруем те регионы в которых приток внутрених туристов есть хотябы в пяти месяцах

Итог: В результате проведенных иследование по пакупательской способности, показателя объема безналичных платежей и притока внутрених туристов, можно выделить четыре ледирующих по этим показателям региона:

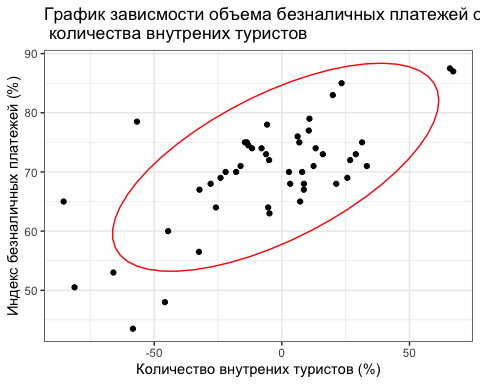
Калининградская область,

Камчатский край,

Кировская область,

Нижегородская область. Показания вышеупамянутых показателей следующие: Показатель покупательской способности. В целом показатель покупателской способности показывает одинаковеое поведение на протежении всего года для всех регионов. Неплоха покупательская активность в начале года в срденем 65-70%, далее идет резкий спад в апреле и мая до 45-50% и постепеный подъем. Максимум наблюдаем в сентябре месяце 75-85% и до конца года держится на высоком уровне. Показатель объема безналичных платежей. Объем безналичных платежей на протежение всего года колеблиться примерно в пределах 55-60%. Без каких-либо скачков для всех регионов. Тоесть чуть больше половины людей пользуется услугами безналичных платежей. Количество вутрених туристов. Количество туристов по сравнению с прошлым годом сильно колеблится на протежение всего года, но тенденция по всем регионом совпадает. Год начинается с увеличение потока внутрених туристов в пределах 5-25%. потом идет на спад и достигает своего миниму апрельмай - июнь, спад может достигать до 80%. И конец лета начало весны наблюдаем повышения потока туристов до 65%. И далее до конца года спад. Можно говорить что указанные регионы в январе, феврале, марте, августе и сентябре посещают наибольшее количество туристов по сравнению с предыдущим годом.

## Поиск взаимосвязи

Построим график зависимости индекса безналичных платежей от колличества внутрених туристов 

Из графика видно, что связь имеется и она прямопропорциональная, тоесть при увеличение количества внутрених туристов идет увеличение объема безналичных платежей.

Посмотрим на коэфициент корреляции Пирсона и проверим значим он или нет. Уровень значимости возмем 5%. И сформулируем нулевую гипотезу, что коэффициент Пирсона равен 0, связи в данных нет. Алтернативная гипотеза будет следующая: связь между между жанных есть и эта связь значимма.

##   
## Pearson's product-moment correlation  
##   
## data: df\_final$Количество.ВТ and df\_final$Индекс.ПА  
## t = 6.5364, df = 46, p-value = 4.53e-08  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5104371 0.8169804  
## sample estimates:  
## cor   
## 0.6939293

Коэффициент корреляции Пирсона равен 0.69. В свою очередь p-value = 4.53е-08, это очень маленькое значение, оно намного меньше 5%. Значит вероятность допустить ошибку отвергнув нелевую гипотезу меньше 5%. Можем отвергнуть нулевую гипотезу. Между данными есть связь и она значима, к томуже коэффициент корреляции положителен это значит, что связь прямопропорциональная.

Предположение Михаила верно, в регионах которые посещают больше туристов доля безналичных платежей выше. Это видно из графика рассеивания и корряляционного теста.

Проверим предположение, что чем больше люди берут кредиты, то тем больше число ипотечных сделок. Для начала переведем данные в количественную шкалу и создадим отдельный датафрейм. Возьмем только начальные значения и переведм их в числовые данные.

##   
## Pearson's product-moment correlation  
##   
## data: df\_ipot\_cred$`Всего одобренных заявок` and df\_ipot\_cred$`Всего ипотечных сделок`  
## t = 7.5654, df = 46, p-value = 1.302e-09  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5840113 0.8490885  
## sample estimates:  
## cor   
## 0.7445903

Проверим предположение, есть ли связь между долей безналичных платежей и количеством поданых заявок онлайн.

##   
## Pearson's product-moment correlation  
##   
## data: df\_final$Индекс.БП and df\_final$`Онлайн-заявки`  
## t = 2.3862, df = 46, p-value = 0.0212  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.0527159 0.5629307  
## sample estimates:  
## cor   
## 0.3318796

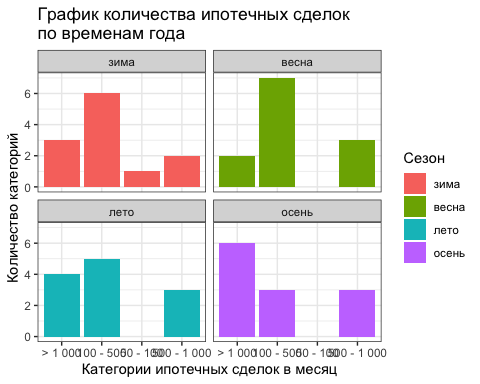
Выводы по пункту 4 и 5.

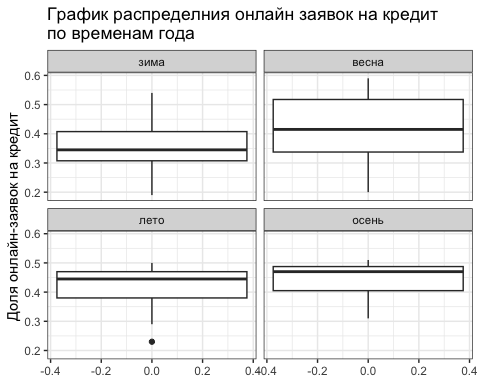
связь между количеством кредитов и числом ипотечных сделок. Сформулируем нулевую гипотезу: связи между парметрами нет, коэффициент корреляции не значим. Альтернативная гипотеза: связь между параметрами есть. Уровень значимости 5%. В результате теста получаем, что коэффициент корреляции Пирсана равен 0.75 - связь сильная и положительная (связь прямопроаорциональная). p-value равна почти нулю, значит нулевую гипотезу можно отбросить, тоесть связь есть. При увеличение количества потребительских кредитов, идет увеличение количество ипотечных сделок.

связь между индексом безнаоичных платежей и количеством онлайн заявок на кредит. Нулевая гипотазе так же как и впредыдущем примере, связи нет между параметрами. Альтернативная шипотеза, связь есть. Уровень значимости возьмем 5%. После проведения теста получаем коэффициент корреляции Пирсона равный 0.33 - связь слабая и положительная (связь прямопропорциональная). p-value равен 0.021, всеровно меньше 5%, это говорит о том, что нулевую гипотезк можно отбросить, связь есть хотя слабая.

Добавим в датафрайм столбец Сезон.

## [1] "зима" "зима" "весна" "весна" "весна" "лето" "лето" "лето" "осень"  
## [10] "осень" "осень" "зима" "зима" "зима" "весна" "весна" "весна" "лето"   
## [19] "лето" "лето" "осень" "осень" "осень" "зима" "зима" "зима" "весна"  
## [28] "весна" "весна" "лето" "лето" "лето" "осень" "осень" "осень" "зима"   
## [37] "зима" "зима" "весна" "весна" "весна" "лето" "лето" "лето" "осень"  
## [46] "осень" "осень" "зима"

Построим столбчатую диаграмму для каждой категории ипотчных сделок в зависимости от времени года.  Из графика видно, что минималльное количество ипотечных сделок приходиться на зиму, а максимальное на осень.

Построим график boxplot для распределения доли онлайн заявок в зависимости от времени года. 

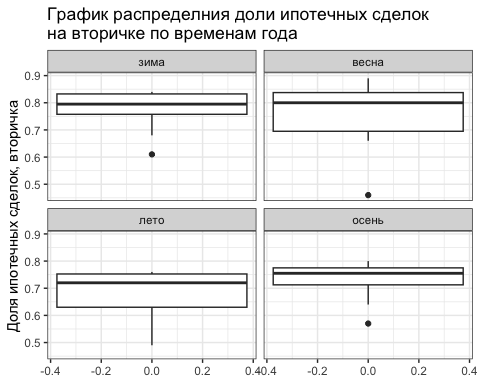
Что бы понять из данного графика в какое время года совершается наибольшее количество онлайн-заявок необходимо сравнить их медианы. Наибольшая медиана приходиться на осень и минималльная на зиму. Таким образом наибольшое количество онлайн заявок приходиться на осень, а минимальное на зиму. Увеличения онлайн заявок осенью, объесняется тем, что люди набирают кредитов в предверии нового года (на подарки или, что-то для себя, дома) сервисы сейчас очень распространены и пользоваться ими очень удобно. Касательно зимнего периода, предположу, что согласно графику движения туристов, в зимний период было увелечения туристов соответственно большинство людей поехали отдыхать. Оформление туристических путевок происходит в офисах компаний, где очень часто находяться представители банков, соответственно оформление кредита происходит в офлайн режиме.

Проверим гипотезу, есть ли связь между временем года и долей онлайн-заявок на кредит. Так как в сравниваемых данных присутствуют категориальные данные воспользуемся тестом хи-квадрат. Сформулируем нулевую гипотезу: зависимости между онлайн-заявками на кредит и временем года нет. Алтернативная гипотеза: зависимость есть. Уровень значимости 5%.

## Warning in chisq.test(df\_final$`Онлайн-заявки`, df\_final$Сезон): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: df\_final$`Онлайн-заявки` and df\_final$Сезон  
## X-squared = 88, df = 78, p-value = 0.2056

По результатом теста получили p-value = 0.21, что больше уровня значимости (5%), значит вероятность допустить ошибку отвергнув нулевую гипотезу более уровня значимости. Соответственно нет оснований отвергнуть нулевую гипотезу и зависимости между временем года и долей онлайн-заявками на кредит нет.

Построим график boxplot для распределения доли ипотечных чделок на вторичке в зависимости от сезона 

Что бы понять в какое время года ипотечных сделок на вторичное жилье было совершено больше на графике boxplot, необходимо сравнить их медианы. Согласно графику, время года в котором было совершено наибольшее количество сделок это зима и весна 2020 года, а минимальное это лето. Возможно это связанно с тем, что в летний период было сдано много нового жилья и люди предпочли брать ипотеки на новое жилье.

Проверим гипотезу, есть ли связь между временем года и долей ипотечных сделок на вторичное жилье. Также воспользуемся тестом хи-квадрат. Сформулируем нулевую и альтернативную гипотезу. Нулевая гипотеза: связи между долей ипотечный сделок на вторичном рынке и временем года нет. Альтернативная гтпотеза: связь есть. Уровень значимости 5%.

## Warning in chisq.test(df\_final$`Доля сделок, вторичка`, : Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: df\_final$`Доля сделок, вторичка` and df\_final$Сезон  
## X-squared = 78.667, df = 72, p-value = 0.2761

По резултам теста p-value = 0.27. Это значение больше уровня значимости. Вероятность дупустить ошибку отвергнув нулевую гипотезу более уровня значимости. Основания отвергнуть нулевую гипотезу нет, соответственно связи между долей ипотечных сделок на вторичке и временем года нет.

Подводя итоги о проделаных изысканий относительно взятых людьми кредитов и ипотек за 2020 год относительно сезона, можно сделать следующие выводы:

максимальное количество ипотечных сделок приходилось на осень, а минимальное на зиму.

из них максимаоьное количество сделок на вторичном рвнке было сделано зимой и весной.

что касается кредитов, то максимальное количество онлайн заявок на кредит за 2020 год было сделано осенью, а минимальное зимой.

все эти данные актуальны на 2020 год, проведенные тесты показали, что каких-то взаимосвязий между ними и временем года нет. Возможно, что бы определить взаимосвязи необходимо больше данных (например взять набор данных за несколько лет) и произвести долполнительные изыскания и тесты.