Разведывательный анализ данных

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

Загрузка данных

1. 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)

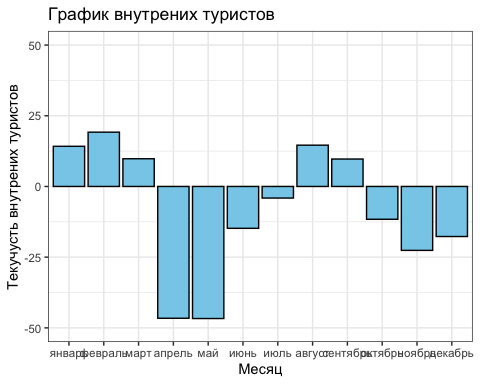
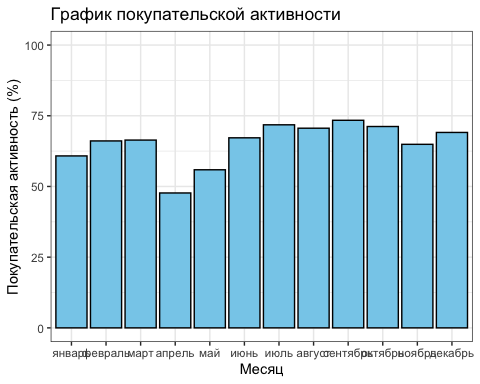
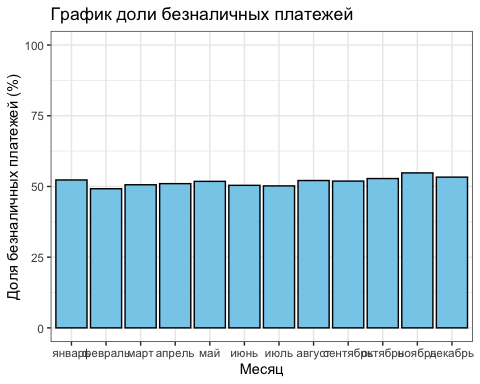
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 34.00 61.00 67.00 65.43 71.00 88.00

summary(df$Количество.ВТ)

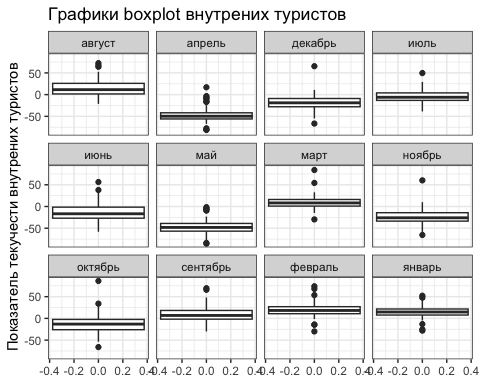
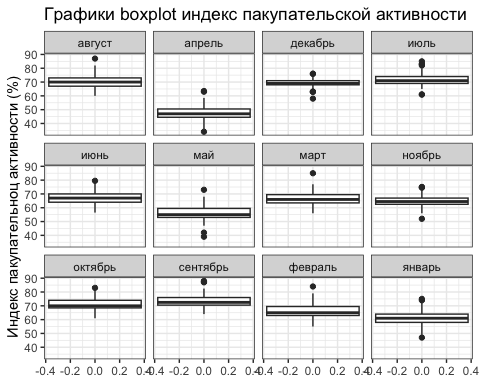
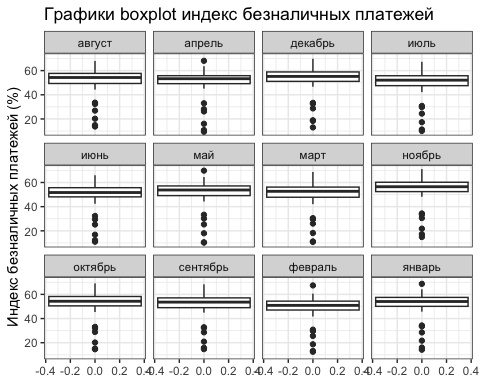
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -85.470 -28.052 -5.125 -8.061 11.162 86.140

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.

1. 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 покупательской 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 внутрених туристов. Так же конец лето и начало осень идет увеличения притока внутренних туристов.

3 .Построим графики boxplot для показателей СберИндекс.



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

Регион

Год

Месяц

Квартал

Число абонентов

Среднемесячная з.п.

Уровень безработицы

Индекс.БП

Индекс.ПА

Количество.ВТ

Всего одобренных заявок

Онлайн-заявки

Офлайн-заявки

Всего ипотечных сделок

Доля сделок, первичка

Доля сделок, вторичка

астраханская область

2020

апрель

квартал 2

745

35800.4

7.9

48.6

45.0

-67.45

100 - 500

0.55

0.45

100 - 500

0.05

0.94

белгородская область

2020

сентябрь

квартал 3

15804

36500.8

5.2

52.3

87.0

6.54

1 000 - 5 000

0.35

0.65

500 - 1 000

0.12

0.85

вологодская область

2020

апрель

квартал 2

2230

40288.4

6.5

58.1

44.0

-54.95

500 - 1 000

0.39

0.61

100 - 500

0.12

0.86

воронежская область

2020

сентябрь

квартал 3

5901

35744.5

4.6

54.3

88.0

13.60

1 000 - 5 000

0.30

0.70

> 1 000

0.35

0.63

ивановская область

2020

апрель

квартал 2

2518

26770.2

5.9

48.9

42.5

-51.80

100 - 500

0.53

0.47

100 - 500

0.05

0.90

кабардино-балкарская республика

2020

апрель

квартал 2

60

28005.8

16.0

26.2

43.5

-50.34

100 - 500

0.43

0.57

100 - 500

0.08

0.87

калининградская область

2020

август

квартал 3

828

36102.0

7.5

59.0

87.0

67.13

1 000 - 5 000

0.44

0.56

500 - 1 000

0.44

0.54

калининградская область

2020

сентябрь

квартал 3

828

36257.7

7.4

58.7

87.5

65.84

1 000 - 5 000

0.43

0.57

500 - 1 000

0.39

0.57

костромская область

2020

апрель

квартал 2

30

28995.0

6.0

55.1

43.0

-43.83

100 - 500

0.36

0.64

100 - 500

0.10

0.89

краснодарский край

2020

апрель

квартал 2

10206

36884.8

6.4

49.8

43.0

-67.22

1 000 - 5 000

0.50

0.50

> 1 000

0.19

0.78

курская область

2020

апрель

квартал 2

1771

34132.5

5.8

48.8

44.0

-35.42

100 - 500

0.33

0.68

100 - 500

0.21

0.77

липецкая область

2020

апрель

квартал 2

5066

34953.2

4.5

50.5

43.5

-32.83

100 - 500

0.42

0.58

100 - 500

0.14

0.84

москва

2020

апрель

квартал 2

41873

101551.2

2.0

58.1

34.0

-77.48

1 000 - 5 000

0.74

0.26

> 1 000

0.24

0.72

москва

2020

май

квартал 2

41873

91823.4

2.3

58.5

42.0

-71.84

5 000 - 10 000

0.71

0.29

> 1 000

0.53

0.45

московская область

2020

апрель

квартал 2

8650

52093.9

3.8

53.6

39.5

-64.36

1 000 - 5 000

0.63

0.37

> 1 000

0.15

0.82

нижегородская область

2020

апрель

квартал 2

832

35818.6

4.7

55.1

43.5

-58.34

1 000 - 5 000

0.54

0.46

> 1 000

0.08

0.89

пермский край

2020

апрель

квартал 2

3207

39337.5

6.7

58.0

41.5

-55.99

1 000 - 5 000

0.50

0.50

> 1 000

0.13

0.83

псковская область

2020

апрель

квартал 2

30

29632.8

6.5

53.9

44.5

-44.22

100 - 500

0.48

0.52

100 - 500

0.15

0.82

республика адыгея

2020

апрель

квартал 2

27

30243.4

8.7

45.7

44.5

-42.44

100 - 500

0.43

0.57

50 - 100

0.10

0.88

республика башкортостан

2020

апрель

квартал 2

825

37123.9

6.2

55.4

44.5

-52.04

1 000 - 5 000

0.41

0.59

> 1 000

0.14

0.84

республика бурятия

2020

апрель

квартал 2

374

39835.5

10.9

54.6

44.5

-41.79

100 - 500

0.61

0.39

100 - 500

0.21

0.78

республика дагестан

2020

апрель

квартал 2

77

27905.8

17.4

16.0

36.0

-50.37

100 - 500

0.53

0.47

100 - 500

0.03

0.94

республика дагестан

2020

май

квартал 2

77

30448.8

16.9

18.0

39.0

-59.21

100 - 500

0.50

0.50

100 - 500

0.14

0.83

республика калмыкия

2020

апрель

квартал 2

391

29739.7

10.1

48.3

41.5

-53.50

100 - 500

0.66

0.34

50 - 100

0.06

0.93

республика карачаево-черкессия

2020

апрель

квартал 2

1043

27986.6

16.4

33.0

44.5

-44.21

100 - 500

0.50

0.50

50 - 100

0.04

0.95

республика карелия

2020

апрель

квартал 2

41873

44129.4

9.4

62.8

40.5

-55.21

100 - 500

0.52

0.48

100 - 500

0.29

0.68

республика коми

2020

апрель

квартал 2

188

56629.3

8.3

62.8

41.5

-54.43

500 - 1 000

0.58

0.42

100 - 500

0.10

0.89

республика марий эл

2020

апрель

квартал 2

87

30360.5

7.4

54.5

44.5

-46.69

100 - 500

0.34

0.66

100 - 500

0.13

0.84

республика северная осетия-алания

2020

апрель

квартал 2

9

28801.4

17.9

28.1

45.5

-62.15

100 - 500

0.48

0.52

100 - 500

0.13

0.79

республика татарстан

2020

апрель

квартал 2

1086

37078.3

3.9

51.3

36.0

-67.45

1 000 - 5 000

0.42

0.58

> 1 000

0.13

0.83

ростовская область

2020

апрель

квартал 2

7352

33201.7

5.2

47.8

41.5

-46.87

1 000 - 5 000

0.46

0.54

500 - 1 000

0.14

0.82

санкт-петербург

2020

апрель

квартал 2

6975

62297.6

2.7

58.2

38.0

-80.89

1 000 - 5 000

0.51

0.49

> 1 000

0.26

0.70

свердловская область

2020

апрель

квартал 2

8591

41114.3

6.8

56.2

45.0

-61.38

1 000 - 5 000

0.48

0.52

> 1 000

0.13

0.83

ульяновская область

2020

апрель

квартал 2

469

31324.8

5.5

47.0

45.0

-47.88

500 - 1 000

0.40

0.60

100 - 500

0.19

0.79

чувашская республика

2020

апрель

квартал 2

2831

29994.3

7.5

49.8

43.0

-53.34

500 - 1 000

0.38

0.62

100 - 500

0.26

0.71

ямало-ненецкий ао

2020

апрель

квартал 2

2064

111954.9

2.4

55.8

42.5

-41.87

100 - 500

0.60

0.40

100 - 500

0.13

0.86

## # 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 нетипичными значениями можно предположить следующее: - В датафрейме с индексом безналичных платежей, видим, что к ним относятся большинство регионов Кавказа (причем в выборку попали регионы полностью по всем месяцам) и показатель всегда ниже первого 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 приходятся 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 не являются выбросами удалять их также не стоит, а обусловлено спадом покупательной активности. То же самое и со значениями превышающие третий квантиль, выпадают на сентябрь месяц, что подтверждает график покупательской активности. - Значения в датафрейме количества внутренних туристов, тоже не стоит относить к выбросам, так как значения больше третьего квантиля и они приходятся на те месяца когда идет рост движения внутренних туристов, что подтверждает график количества внутренних туристов. Если говорить в целом нетипичные значения отличаются примерно на 1-5%, на мой взгляд не стоит удалять их. Они не повлияют на модель, а из-за их удаления можем потерять много полезной информации.

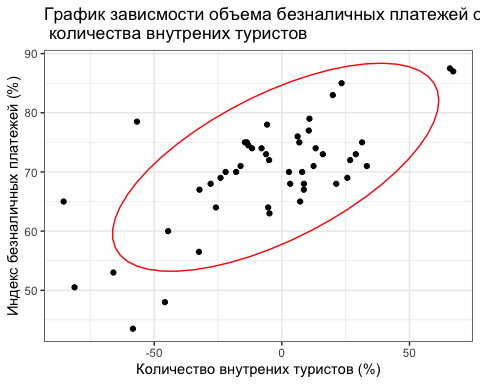
1. Выберем топ 30 регионов лидирующих по индексу покупательской способности
2. Выберем топ 30 регионов лидирующих по индексу безналичных платежей.
3. Выберем те регионы которые лидируют по индексу покупательской способности и индексу безналичных платежей.
4. Выберем строки из датафрейма df которые соответствуют выборки лидирующих регионов

9,10. Отфильтруем те регионы в которых приток внутренних туристов есть хотя бы в пяти месяцах

1. Итог: В результате проведенных исследование по покупательской способности, показателя объема безналичных платежей и притока внутренних туристов, можно выделить четыре лидирующих по этим показателям региона:

* Калининградская область,
* Камчатский край,
* Кировская область,
* Нижегородская область. Показания вышеупомянутых показателей следующие: Показатель покупательской способности. В целом показатель покупательской способности показывает одинаковое поведение на протяжении всего года для всех регионов. Неплоха покупательская активность в начале года в среднем 65-70%, далее идет резкий спад в апреле и мая до 45-50% и постепенный подъем. Максимум наблюдаем в сентябре месяце 75-85% и до конца года держится на высоком уровне. Показатель объема безналичных платежей. Объем безналичных платежей на протяжение всего года колеблется примерно в пределах 55-60%. Без каких-либо скачков для всех регионов. То есть чуть больше половины людей пользуется услугами безналичных платежей. Количество внутренних туристов. Количество туристов по сравнению с прошлым годом сильно колеблется на протяжение всего года, но тенденция по всем регионом совпадает. Год начинается с увеличение потока внутренних туристов в пределах 5-25%. потом идет на спад и достигает своего минимум апрель - 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

1. Построим график зависимости индекса безналичных платежей от количества внутренних туристов 

Из графика видно, что связь имеется и она прямопропорциональная, то есть при увеличение количества внутренних туристов идет увеличение объема безналичных платежей.

1. Посмотрим на коэффициент корреляции Пирсона и проверим значим он или нет. Уровень значимости возьмем 5%. И сформулируем нулевую гипотезу, что коэффициент Пирсона равен 0, связи в данных нет. Альтернативная гипотеза будет следующая: связь между данных есть и эта связь значима.

##   
## 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 нулевую гипотезу меньше 5%. Можем отвергнуть нулевую гипотезу. Между данными есть связь и она значима, к тому же коэффициент корреляции положителен это значит, что связь прямопропорциональная.

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

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

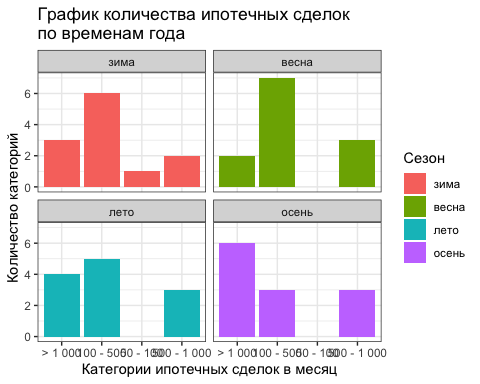
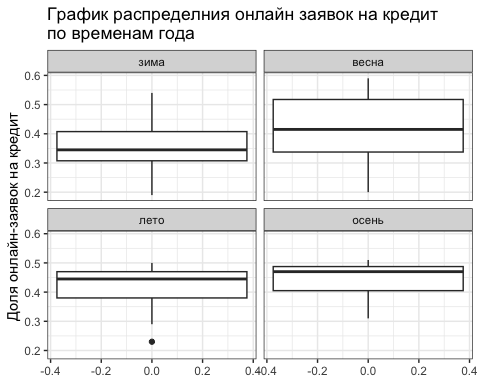
##   
## Pearson's product-moment correlation  
##   
## data: df\_final$Индекс.БП and df\_final$`Онлайн-заявки`  
## t = 2.3862, df = 46, p-value = 0.0212  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.0527159 0.5629307  
## sample estimates: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

1. Выводы по пункту 4 и 5.

* связь между количеством кредитов и числом ипотечных сделок. Сформулируем нулевую гипотезу: связи между варметрами 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, то есть связь есть. При увеличение количества потребительских кредитов, идет увеличение количество ипотечных сделок.
* связь между индексом безналичных платежей и количеством онлайн заявок на кредит. Нулевая гипотезе так же как и в предыдущем примере, связи нет между параметрами. Альтернативная гипотеза, 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 гипотеза можно отбросить, связь есть хотя слабая.

1. 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"

1. Построим столбчатую диаграмму для каждой категории ипотечных сделок в зависимости от времени года.  Из графика видно, что минимальное количество ипотечных сделок приходиться на зиму, а максимальное на осень.
2. Построим график boxplot для распределения доли онлайн заявок в зависимости от времени года. 

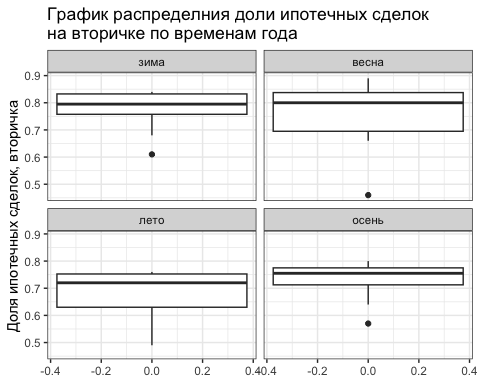
Что бы понять из данного графика в какое время года совершается наибольшее количество онлайн-заявок необходимо сравнить их медианы. Наибольшая медиана приходиться на осень и минимальная на зиму. Таким образом наибольшие количество онлайн заявок приходиться на осень, а минимальное на зиму. Увеличения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

1. 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. Альтернативная 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.

1. Построим график boxplot для распределения доли ипотечных сделок 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.

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

1. Подводя итоги о проделанных изысканий относительно взятых людьми кредитов и ипотеки за 2020 год относительно сезона, можно сделать следующие выводы:

* максимальное количество ипотечных сделок приходилось на осень, а минимальное на зиму.
* из них максимальное количество сделок на вторичном рынке 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 взаимосвязей между ними и временем года нет. Возможно, что бы определить взаимосвязи необходимо больше данных (например взять набор данных за несколько лет) и произвести дополнительные изыскания и тесты.