Презентация дипломной работы

Дипломная работа по программе «Аналитика данных для начинающих»

Власов А.А.
Группа: DA-CP-3

Введение

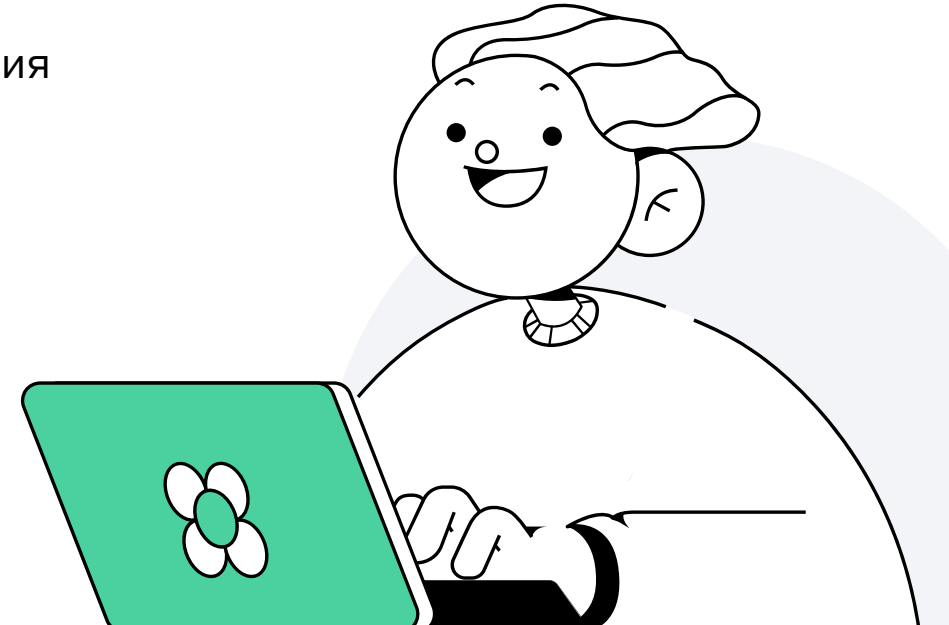
Дипломная работа на тему **“Анализ данных крупного агрегатора такси (поиск инсайтов, проверка гипотезы, составление рекомендаций стейкхолдерам)”**.

Работа выполнена обучающимся по программе “Аналитика данных для начинающих” **Власовым Александром Александровичем**

В презентации представлены итоги исследования согласно имеющимся материалам для исследования

Структура презентации

1. Цель дипломной работы
2. Задачи дипломной работы
3. Описание данных
4. Описание проведенного исследования
5. Результаты проверки гипотез
6. Выводы и рекомендации



Цель дипломной работы

1

Цель дипломной работы

Собрать все полученные знания воедино

Задачи дипломной работы

2

Задачи дипломной работы

- **Провести исследование данных**
- **Проверить гипотезу о поведенческом предпочтении пользователей к определённому классу такси**
- **Сделать выводы согласно полученным результатам, чтобы помочь компании стать лучшей на рынке**
- **Презентовать результаты**

Описание данных

3

Описание данных

Данные представлены в файле формата csv

Описание датасета:

Датасет содержит информацию о поездках на такси одного из известных агрегаторов:

- order_gk — идентификатор поездки;
- weekday_key — день недели совершения поездки;
- hour_key — час совершения поездки;
- distance_km — дистанция поездки;
- offer_class_group — класс поездки (Economy, Comfort, Premium, Delivery);
- driver_response — статус поездки (1 — поездка совершена; 0 — отмена)

Описание проведённого исследования

4

Описание проведённого исследования

- Загрузка файла dip_hw_x_taxi.csv в pandas dataframe
- Расчёт основных описательных статистик
- Вычисление количества значений по классам такси
- Визуализация распределений дистанций поездок
- Визуализация распределений дистанций поездок, по каждому классу такси в отдельности, а также определение типов распределений для эконом-класса и комфорт-класса
- Формирование выборки по эконом и комфорт-классам. Визуализация пересечений интервалов дистанций этих классов
- Проведение стат. тест, проверка гипотезы, что дистанции поездок в комфорт-классе отличаются от дистанций поездок эконом-класса (уровень значимости = 5%)
- Вычисление средних выборок исследуемых классов, определение, с учётом полученного результата проверки, гипотезы – какой из классов предпочитают на дальние поездки (эконом или комфорт)

Загрузка файла dip_hw_x_taxi.csv в pandas dataframe

```
taxi_info = pd.read_csv('.../dip_hw_x_taxi.csv')
```

Расчёт основных описательных статистик

	order_gk	weekday_key	hour_key	distance_km	driver_response
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	462807.384000	4.044000	11.608500	26.275848	0.587000
std	270556.211847	2.007256	6.918465	18.876336	0.492496
min	412.000000	1.000000	0.000000	0.024000	0.000000
25%	227126.500000	2.000000	6.000000	12.177000	0.000000
50%	458637.000000	4.000000	12.000000	22.828500	1.000000
75%	709420.250000	6.000000	18.000000	36.434250	1.000000
max	919196.000000	7.000000	23.000000	138.950000	1.000000

Расчёт основных описательных статистик

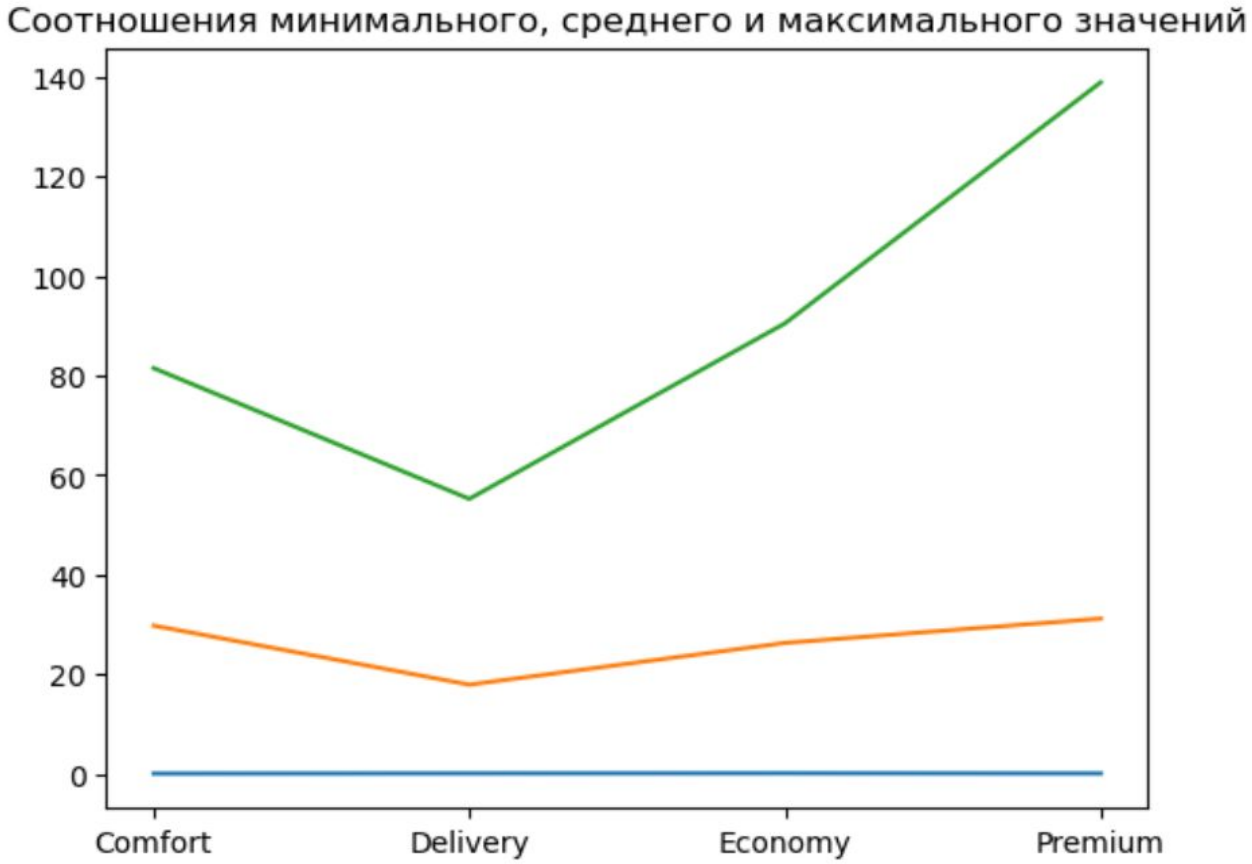
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   order_gk            2000 non-null   int64
 1   weekday_key         2000 non-null   int64
 2   hour_key            2000 non-null   int64
 3   distance_km         2000 non-null   float64
 4   offer_class_group   2000 non-null   object
 5   driver_response     2000 non-null   int64
dtypes: float64(1), int64(4), object(1)
memory usage: 93.9+ KB
```

Вычисление количества значений по классам такси

:

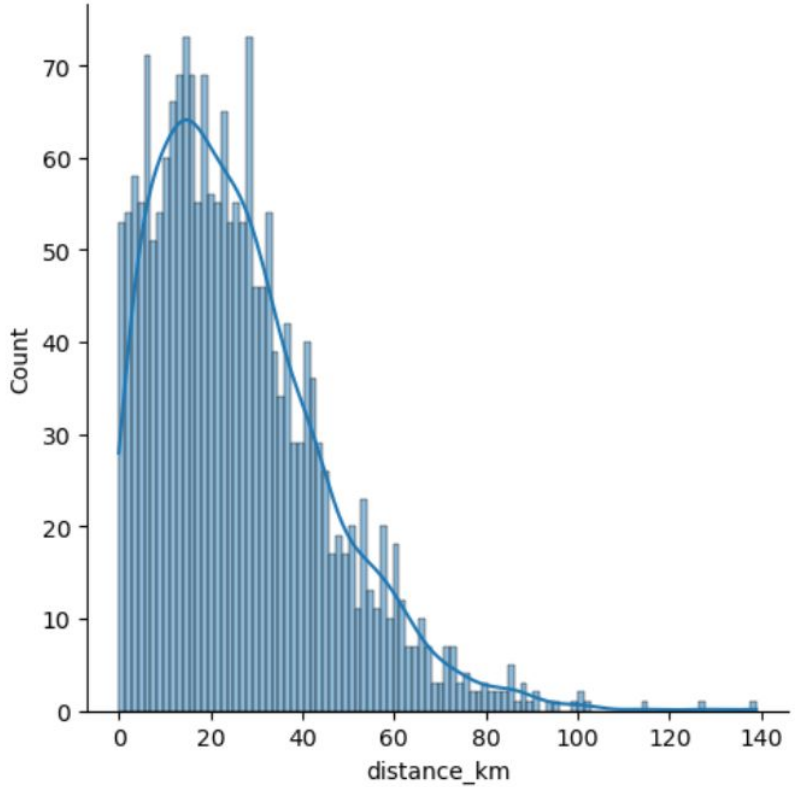
	offer_class_group	order_gk	weekday_key	hour_key	distance_km	driver_response
0	Comfort	500	500	500	500	500
1	Delivery	500	500	500	500	500
2	Economy	500	500	500	500	500
3	Premium	500	500	500	500	500

Визуализация распределений дистанций поездок



Визуализация распределений дистанций поездок

Соотношение расстояния поездки и количества соответствующих дистанций

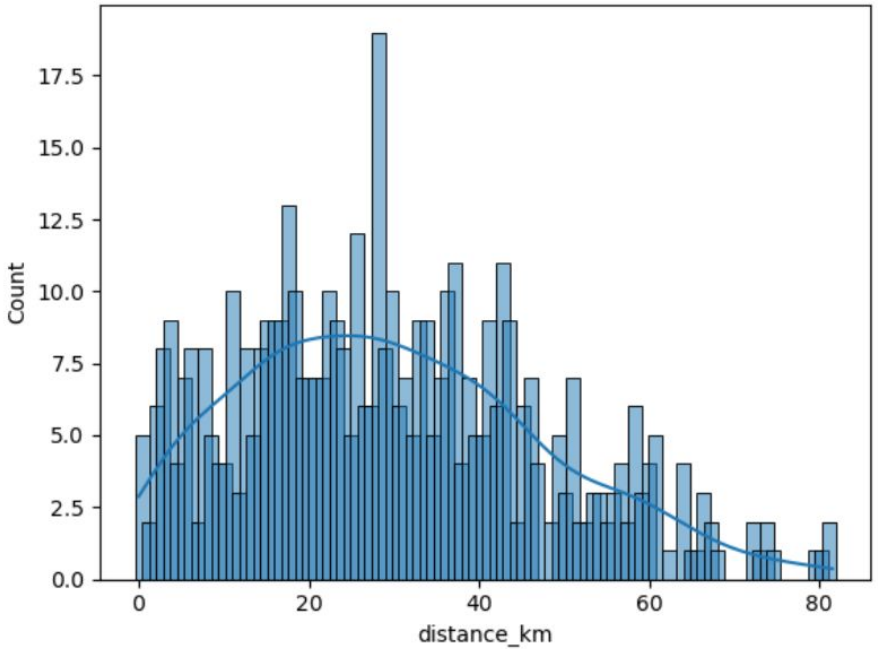


Визуализация распределений дистанций поездок, по каждому классу такси в отдельности, а также определение типов распределений для эконом-класса и комфорт-класса

- taxi_info_comfort = taxi_info.where(taxi_info.offer_class_group == 'Comfort')
- taxi_info_economy = taxi_info.where(taxi_info.offer_class_group == 'Economy')
- taxi_info_premium = taxi_info.where(taxi_info.offer_class_group == 'Premium')
- taxi_info_delivery = taxi_info.where(taxi_info.offer_class_group == 'Delivery')

Визуализация распределений дистанций поездок

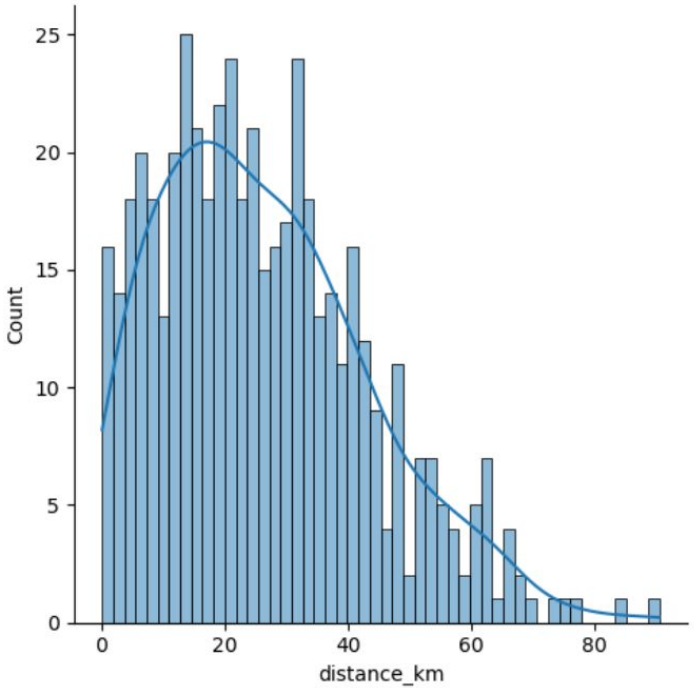
Распределение дистанций поездок по классу Комфорт



Распределение "нормальное" или "с изолированным пиком"

Визуализация распределений дистанций поездок

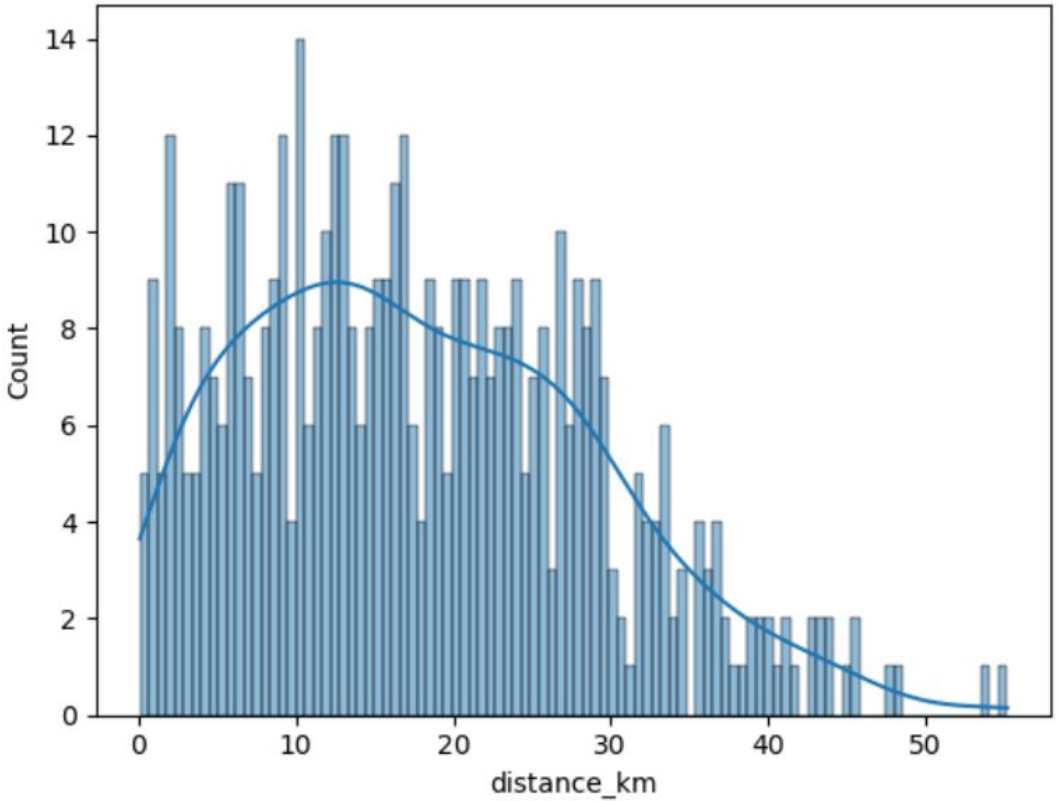
Распределение дистанций поездок по классу Эконом



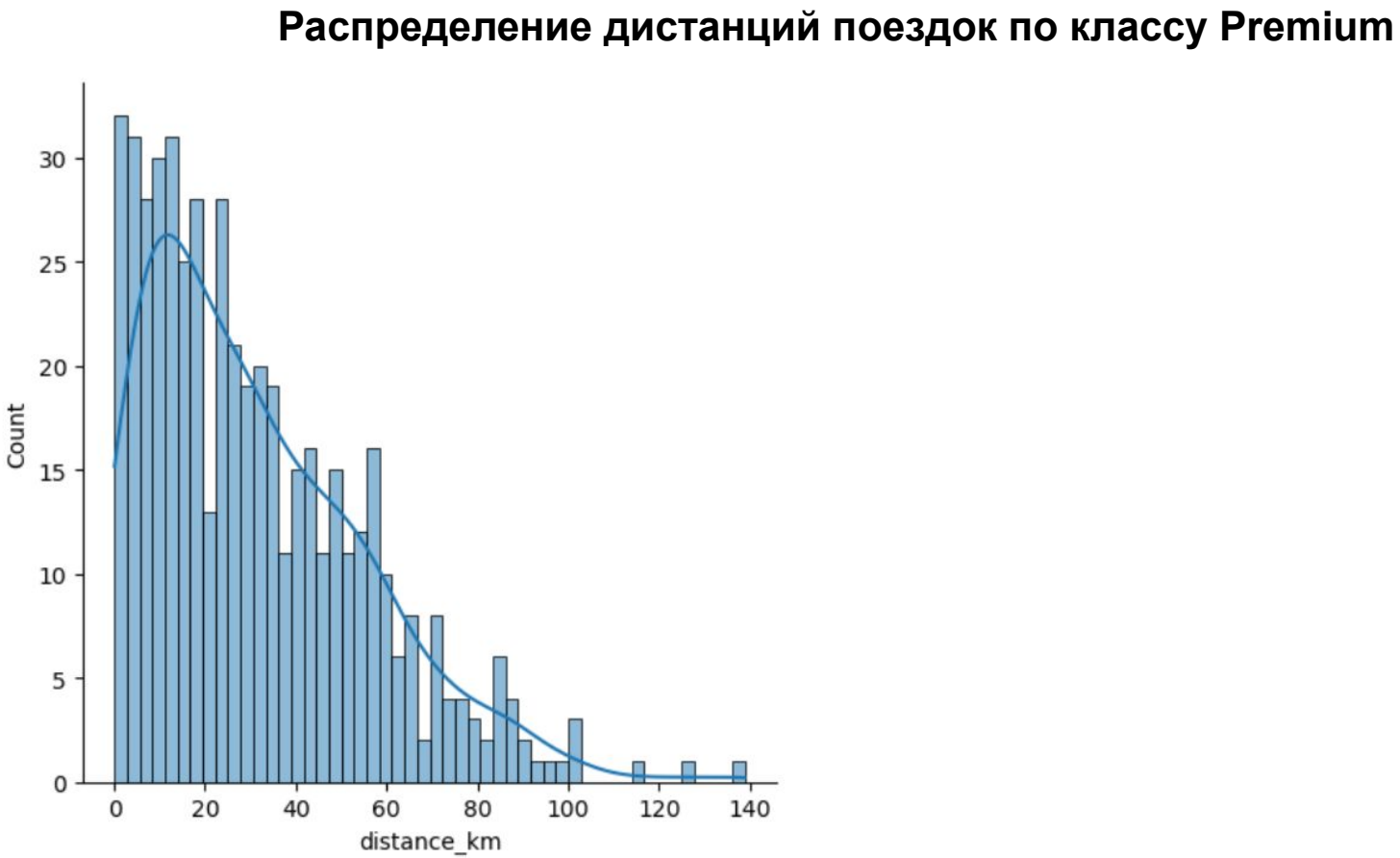
"Положительное или отрицательное скошенное распределение"

Визуализация распределений дистанций поездок

Распределение дистанций поездок по классу Delivery



Визуализация распределений дистанций поездок



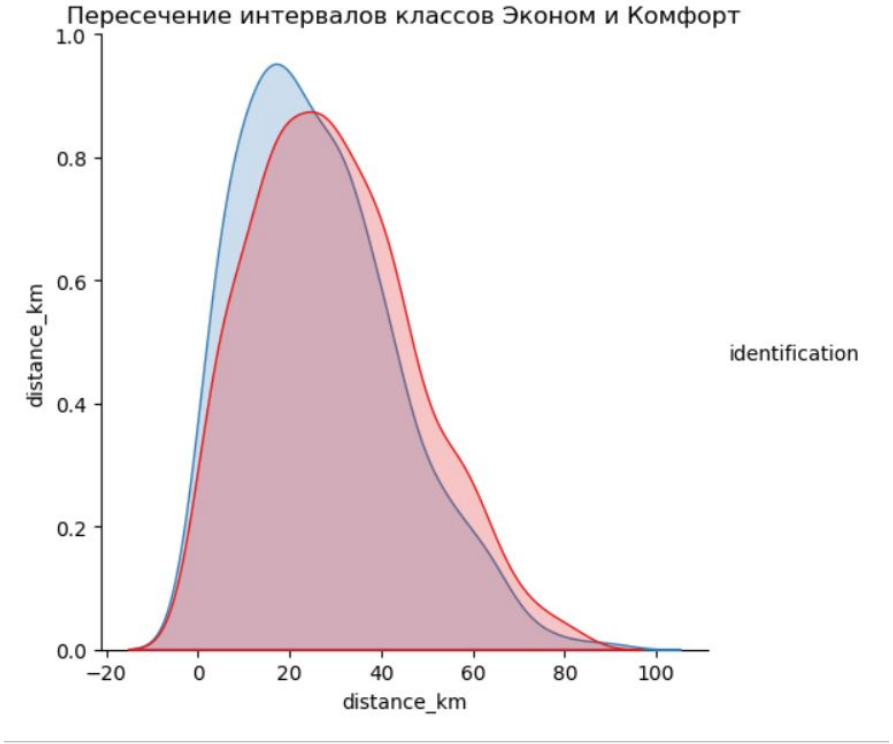
Формирование выборки по эконом и комфорт-классам

taxi_info_eccom = pd.concat([taxi_info_comfort,taxi_info_economy], axis=0, ignore_index=True)

	order_gk	weekday_key	hour_key	distance_km	offer_class_group	driver_response	identification
0	289990.0	5.0	0.0	0.024	Comfort	0.0	1.0
1	242330.0	5.0	0.0	0.144	Comfort	0.0	1.0
2	798867.0	5.0	14.0	0.292	Comfort	1.0	1.0
3	577776.0	3.0	21.0	0.422	Comfort	1.0	1.0
4	68734.0	5.0	16.0	0.609	Comfort	1.0	1.0
...	...	...	...	...	...	...	...
995	266312.0	6.0	14.0	73.527	Economy	0.0	2.0
996	407927.0	6.0	0.0	75.213	Economy	0.0	2.0
997	369139.0	6.0	13.0	77.648	Economy	0.0	2.0
998	3845.0	3.0	7.0	83.830	Economy	0.0	2.0
999	77493.0	7.0	9.0	90.517	Economy	0.0	2.0

1000 rows × 7 columns

Визуализация пересечений интервалов дистанций эконом и комфорт-классов



Одинаковое количество заказов такси эконом и комфорт класса распределено в интервале от 26 до 28 километров

Проведение стат. тест, проверка гипотезы, что дистанции поездок в комфорт-классе отличаются от дистанций поездок эконом-класса (уровень значимости = 5%)

Проверка гипотезы на случайной выборке:

X_train, X_test, y_train, y_test = train_test_split(taxi_info_comfort, taxi_info_economy, test_size=0.2)

ttest_ind(X_train['distance_km'], y_train['distance_km'])

Ttest_indResult(statistic=2.920871027493813, pvalue=0.003588957297011715)

Проверка гипотезы на всей выборке:

ttest_ind(taxi_info_comfort['distance_km'], taxi_info_economy['distance_km'])

Ttest_indResult(statistic=3.1706426042389753, pvalue=0.0015673809987275806)

По двум выборкам мы получаем значение p-value меньше допустимого уровня значимости

Вычисление средних выборок исследуемых классов, определение, с учётом полученного результата проверки, гипотезы – какой из классов предпочитают на дальние поездки (эконом или комфорт)

Среднее значение дистанций по классам Комфорт и Эконом:

	offer_class_group	distance_km
0	Comfort	29.758916
1	Economy	26.283098

Вычисление средних выборок исследуемых классов, определение, с учётом полученного результата проверки, гипотезы – какой из классов предпочитают на дальние поездки (эконом или комфорт)

Вычисление медианного значения:

taxi_info_eccom.groupby('offer_class_group').median()

offer_class_group	order_gk	weekday_key	hour_key	distance_km	driver_response	identification
Comfort	475479.0	4.0	12.0	28.0785	0.0	1.0
Economy	455884.5	4.0	12.0	23.7980	0.0	2.0

taxi_info_comfort.where(taxi_info_comfort.distance_km >= 28) 252

taxi_info_economy.where(taxi_info_economy.distance_km >= 28) 215

Количество поездок класса Комфорт с дистанцией выше или равной 28 км - медианное значение, - больше чем количество поездок класса Эконом с этим же диапазоном дистанций. в следствии чего можно утверждать, что такси класса Комфорт пользуется большим спросом при длительных поездках.

Результаты проверки гипотез

5

Результаты проверки гипотез

1. **В результате проведение стат. теста, в рамках которого была проведена проверка гипотезы, что дистанции поездок в комфорт-классе отличаются от дистанций поездок эконом-класса мы получили значение p-value меньше допустимого уровня значимости. По итогам теста мы отвергаем нулевую гипотезу, получаем, что дистанции поездок в комфорт-классе отличаются от дистанций поездок эконом-класса.**
2. **В результате вычисления средних выборок исследуемых классов и определения, с учётом полученного результата проверки, гипотезы – какой из классов предпочитают на дальние поездки (эконом или комфорт) было определено, что такси класса Комфорт пользуется большим спросом при длительных поездках.**

Выводы и рекомендации

6

Выводы и рекомендации

- Исследования проведены в области сферы услуг, проанализирована выборка поездок на такси одного из известных агрегаторов. Были сделаны заключения: класс такси Delivery больше всего используется для коротких поездок; класс такси Premium затрагивает дальние дистанции, но по количеству поездок данного типа такси можно сказать, что большая плотность в области до 50 км.; классы Комфорт и Эконом по произведенным поездкам практически одинаковые, но класс Комфорт на дальние дистанции используется чаще чем класс Эконом.
- Таким образом можно сказать, что все классы такси оправдывают себя и в полной мере выполняют свои функции
- Рекомендации для стейкхолдеров данного исследования. Для выравнивания данных класса Эконом по отношению к классу Комфорт можно порекомендовать проанализировать уровень сервиса и по возможности улучшить некоторые пункты(например это может быть уровень удобства салона). Также для более продуктивной работы можно сразу распределять заказы по классам такси согласно заявленной дистанции поездки и рекомендовать клиенту при заказе определенный класс такси

Спасибо за внимание!