**Итоговая работа по курсу**

**«Аналитик Данных»**

**Исследование популярности ресторанов Европы (анализ данных, составление рекомендаций, построение дашборда)**

**Работу выполнила: Студентка группы DA-12**

**Несмеянова Юлия**

1. **Постановка задачи**

В рамках исследования поставлена задача анализа ресторанов Европы. Попытаемся ответить на следующие вопросы: 1) Что влияет на успешность ресторана? Какими характеристиками должен обладать ресторан, чтобы стать популярным?

Целью ставится получение наглядных результатов, которые могут быть использованы потенциальными рестораторами для развития бизнеса, открытия новых точек.

Итогом работы будет создан дашборд с визуализацией. Актуальность дашборда состоит в том, что он позволит наглядно получить краткую информацию о сфере и использовать ее для реальных бизнес-задач, таких как: принятие решений об открытии новых ресторанов; принятие решении об переустройстве действующих заведений (например, смена формата или смена региональной кухни) и т.д. Результаты исследования могут быть интересны следующим стейкхолдерам:

1. Рестораторы, развивающие свои точки в Европе;
2. Поставщики продукции;
3. Туристы, предпочитающие гастрономический туризм;
4. Специалисты, потенциальные работники сферы
5. **Формулировка гипотез**

Успешность ресторана определяется его оценкой потребителями. Следовательно центральным показателем оценивания ресторана выбрана оценка посетителей.

Сформируем гипотезы для дальнейшей проверки:

1) На оценку посетителей ресторана влияет число отзывов – можно предположить, что рестораны с большим количеством отзывов более популярны, их выбирает большее число потребителей, а следовательно эти рестораны более популярны, вероятно их оценка выше, чем у менее оцениваемых ресторанов;

2) На оценку посетителей ресторана влияет его ценовой критерий – предположим, что более дорогие рестораны имеют оценку ниже, чем рестораны с небольшим ценовым критерием;

3) На оценку посетителей ресторана влияет вид его кухни – предположим, что более популярные кухни оцениваются выше, чем экзотичные, непривычные европейскому вкусу.

Для проверки гипотезы взяты данные с сайта Kaggle ([ссылка на данные](https://www.kaggle.com/damienbeneschi/krakow-ta-restaurans-data-raw))

1. **Предобработка данных**

Исходные данные представляют собой датасет, собранный с популярного туристического сайта TripAdvisor. Данные содержат информацию о ресторанах тридцати одного города Европы. Размер датасета 127527 строк и 10 столбцов. По некоторым столбцам имеются пропуски. Данные содержат следующую информацию:

1. Название ресторана – индивидуальное название ресторана (есть небольшие повторения ресторанов с одинаковым названием в разных городах – 4101 ресторан представлен более чем в одном городе, самые популярные из них – точки фастфудных сетей)
2. Город - название города. В данных содержится информация по тридцати одному крупному европейскому городу. Города, попавшие в выборку – Амстердам, Афины, Барселона, Берлин, Братислава, Брюссель, Будапешт, Копенгаген, Дублин, Эдинбург, Женева, Гамбург, Хельсинки, Краков, Лиссабон, Любляна, Лондон, Люксембург, Лион, Мадрид, Милан, Мюнхен, Опорто, Осло, Париж, Прага, Рим, Стокгольм, Вена, Варшава и Цюрих. Распределение по городам не равное. Так больше всего ресторанов в Лондоне – 18212, а меньше всего в Любляне – всего 501.
3. Стиль кухни – содержит список кухонь, представленных в ресторане.

В данном столбце может содержаться информация о национальном виде кухне – например – итальянская, английская, датская; типе ресторана – например – бар, гриль, паб; особенностях кухни – например – вегетарианская кухня, без глютена и т.д.

1. Рейтинг ресторана – оценка ресторана среди других заведений в рамках одного города. Данная информация может быть использована для исследования в разрезе городов.
2. Оценка ресторана – оценка от 1 до 5. Принимаем данный показатель результативным – именно оценка ресторана определяет его успешность. В дальнейшем будет исследовано как данный показатель зависит от различных факторов.
3. Ценовой критерий– информация о среднем ценовом критерии по ресторану. Категориальная переменная, которая может принимать 3 значения.
4. Число отзывов – число отзывов, оставленных на сайте посетителями ресторана.
5. Отзывы – два случайных отзыва с первой странницы сайта с отзывами. Данная информация также не актуальна в рамках работы.
6. URL-ссылка - ссылка на страницу о ресторане на сайте TripAdvisor.
7. Индивидуальный номер

Для выполнения анализа и проверки сформулированных гипотез данные необходимо подвергнуть предобработке.

Итого данные подверглись следующей обработке:

1. Из датасета удалены столбцы не актуальные в исследовании;
2. Названия ресторанов приведены к одному регистру, удалены строки с дубликатами;
3. Информация о среднем ценовом критерии е в ресторане приведена к числовому виду;
4. Заполнены средним значением по городу пропуски по столбцам с информацией о ценах и числе отзывов. Удалена часть строк с пропусками – 7,1% выборки;
5. Проведена категоризация ресторанов по видам кухни в соответствии с гипотезами: 1. Отмечены рестораны с европейской кухней – предполагаем, что европейская кухня близка европейскому потребителю и следовательно такие рестораны получают более высокую оценку; 2. Отмечены рестораны, предлагающие здоровое питание – веганские, вегетарианские и блюда без глютена – предположим оценку выше, за счет посетителей, предпочитающих правильное питание. 3. Отмечены рестораны, предлагающие алкоголь – предположим более низкую оценку, так как такие рестораны не претендуют на высокую кухню (это бары и пабы). 4. Отмечены заведения фастфуда – предполагаем более низкую оценку таких заведений. 5. Отмечены рестораны с неопределенной кухней – предположим, что такие рестораны имеют более низкие оценки, так как можно предположить, что к таким заведениям относятся заведения непопулярные, имеющие мало отзывов и следовательно не высоко оцениваемые посетителями.

Также имеется рестораны с большим количеством других кухонь, однако они в данной работе не рассматриваются так как либо количество таких ресторанов очень мало по сравнению с выборкой, как например ресторанов с египетской кухней – 62, либо категория кухни ресторана не позволяет выдвинуть обоснованную гипотезу, например сложно сказать будут ли рестораны с азиатской кухней получать более высокие или более низкие оценки.

В результате первичной обработки данных получили обработанный датасет готовый для дальнейшей работы.

1. **Проверка гипотез**

Для проверки гипотез строим матрицу корреляций. В результате получаем следующие значения:

1. Корреляция между оценкой ресторана и его средним ценовым критерием отрицательная – как и предполагалось в гипотезе, чем выше ценовой критерий ресторана, тем ниже его оценка, однако значение коэффициента незначительно - зависимость составляет примерно 2,1%
2. Корреляция между оценкой ресторана и числом отзывов на его положительная, чем больше отзывов, тем выше оценка ресторана – как и предполагалось в гипотезе – коэффициент имеет значение – 3,2%
3. Европейская кухня, здоровое питание и наличие алкоголя положительно влияют на оценку – 4,7%, 14% и 1,3% соответственно. Европейская кухня и здоровое питание подтверждают гипотезу, а алкогольное меню ее опровергает.
4. Фастфуд и неопределенность кухни негативно влияют на оценку ресторана – как и предполагалось в гипотезе, но также с небольшими значениями – 0,9% и 4,2% соответственно.

Используем сводные таблицы что бы посмотреть средние значения. График взаимосвязи ценового критерия а и средней оценки не линеен. Можно наблюдать вначале небольшое снижение, затем множественные колебания и рост. График взаимосвязи числа отзывов и средней оценки и числа отзыва также демонстрирует колебания, однако можно увидеть незначительное увеличение средней оценки при росте числа отзывов.

Характеристики кухонь незначительно влияют на среднюю оценку ресторана. Только критерий наличия здорового питания показывает заметную разницу – так средняя оценка ресторанов, предоставляющих данную опцию -4.14¸ а не предоставляющих - 3.93

Таким образом можно сделать вывод о том, что подтверждены следующие гипотезы:

1) Наличие опции здоровое питание в кухне положительно влияет на среднюю оценку ресторана;

2) Европейская кухня в ресторане, незначительно, но положительно влияет на среднюю оценку ресторана;

3) Число отзывов на ресторан незначительно, но положительно влияет на среднюю оценку ресторана посетителем.

Корреляция между оценкой ресторана и его средним ценовым критериям составляет всего 2,1%. Проверим значимость этого коэффициента с помощью теста. Для этого разделим совокупность данным по ценовому критерию и проведем статистический тест на равенство средних. Сформулируем нулевую гипотезу H0 – средняя оценка для разных ценовых групп равна. H1 – средняя оценка для разных ценовых групп отличается. Установим уровень значимости в 5%. Используем t-тест. Так как значения в одном ресторане не зависит от другого используем тест для независимых выборок. Получаем p-значение равное 0.00000000446. Оно значительно меньше установленного уровня значимости. Отвергаем нулевую гипотезу. Существенных различий в оценке

ресторанов с ценовым критерием выше среднего и ниже среднего нет.

1. **Построение дашборда**

Продолжим исследование рассматривая данные в разрезе городов. Для этого проведем визуальный анализ, так как он хорошо подходит для сравнения показателей по городам.

Результатом будет дашборд, выполненный средствами Power BI.

Гистограмма хорошо показывает неравномерное распределение числа ресторанов по городам. Так более половины всех ресторанов приходится на 7 крупных городов – Лондон, Париж, Мадрид, Барселона, Берлин, Милан и Рим. При этом Лондон и Париж занимают соответственно 14,14% и 12,1 % выборки. Все остальные города представляют менее чем 5% выборки каждый. Особенно незначительный вклад внесли такие города как Любляна, Люксембург, Братислава, Осло, Хельсинки – каждый менее 1%.

В разрезе средней оценке ресторанов по городам самые высокие показатели у Рима – 4,23, Афин – 4,21 и Опорто – 4,15. При это Афины и Опорто также являются одними из самых дешевых городов – 2,04 и 2,01 средний показатель ценового критерия – все это делает эти города привлекательными в рамках соотношение «цена/качество». Самые низкие оценки получают рестораны Мадрида, Милана и Стокгольма – 3,79, 3,81 и 3,87 соответственно. Самым «дешевым» городом оказывается Краков – показатель 1,94. Самые «дорогие» города – Женева, Цюрих и Люксембург.

В разрезе опций – Рим также является городом с наименьшей долей ресторанов с неопределенной кухней – 0,12. Самый высокий процент неопределенности в городе Гамбург.

Наибольшее число ресторанов с алкоголем предлагает Дублин – доля таких заведений в городе 0,23. Меньше всего алкогольных заведений в том же Гамбурге.

Фастфуд популярнее всего в Эдинбурге – доля составила 0,11, наименее популярен в Гамбурге – 0,02.

Эдинбург может предложит большое количество ресторанов со здоровой едой – 0,54 от всех ресторанов города. В Лионе доля заведений со здоровой пищей всего 0,10.

Европейскую кухню в 0,58 от всех ресторанов города предложат в Опорто. Наименьшее число ресторанов с европейской кухней в Милане – 0,09 от всех заведений города.

1. **Вывод**

На основе проведенного исследования можно выдвинуть следующие рекомендации:

1. Для посетителей из Европы и туристов, имеющих своей целью гастрономический туризм:

* Выбирая город для посещения с целью распробовать местные заведения следует отдать свое предпочтение таким городам как – Рим, Лондон, Париж
* Если цель сэкономить, следует обратить внимание на такие города, как – Любляна, Краков, Афины
* Выбирая между ресторанами стоит предпочесть рестораны с местной кухней и особенно рестораны, предлагающие здоровое питание. Также не лишним будет обратить внимание на рестораны с большим количеством отзывов.

1. Для стороны бизнеса:

* При определении кухни заведения лучше остановиться на классической и привычной;
* Здоровое питание популярнее фастфуда в Европе – это тоже следует учесть.