ГУАП КАФЕДРА № 41

ОТЧЕТ

ЗАЩИЩЕН С ОЦЕНКОЙ ПРЕПОДАВАТЕЛЬ

Старший преподаватель В. В. Боженко

должность, уч. степень, звание подпись, дата инициалы, фамилия

# ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №1

ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ

по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ

РАБОТУ ВЫПОЛНИЛ

СТУДЕНТ ГР. № 4117 Д. С. Николаев

подпись, дата инициалы, фамилия

Санкт-Петербург 2023

# **Цель работы**

Осуществить предварительную обработку данных csv-файла, выявить и устранить проблемы в этих данных.

# **Индивидуальное задание**

Вариант 2.

Задание 1: Группировка - CATEGORY и количество поездок каждого типа (по цели маршрута).

Задание 2: Группировка - CATEGORY и количество поездок каждого типа (по цели маршрута). Создать датафрейм. Переименовать столбец с количеством в “сount”. Отсортировать по убыванию столбца “count”.

Задание 3: Сводная таблица (pivot\_table) - средняя количество пройденных миль по каждой цели поездки (PURPOSEroute). Отсортировать по убыванию столбца MILES. Округлить значение до двух знаков.

Задание 4: Сводная таблица (pivot\_table) - средняя количество пройденных миль по каждой цели поездки (PURPOSEroute) - столбцы и каждой категории - строки. Отсортировать по убыванию столбца CATEGORY.

# **Ход работы**

Ссылка на [gitHub](https://github.com/TanukiY/DataAnalysis).

Для начала был загружен датасет через библиотеку Python – pandas. Используем ';' для разделения данных. Код расположен в листинге 1.

Листинг 1 - Код загрузка датасета

import pandas

df = pandas.read\_csv("drivers.csv", sep =';')

Далее были выведены первые 20 строк через метод df.head(20). На рисунке 1 продемонстрирован результат работы метода.

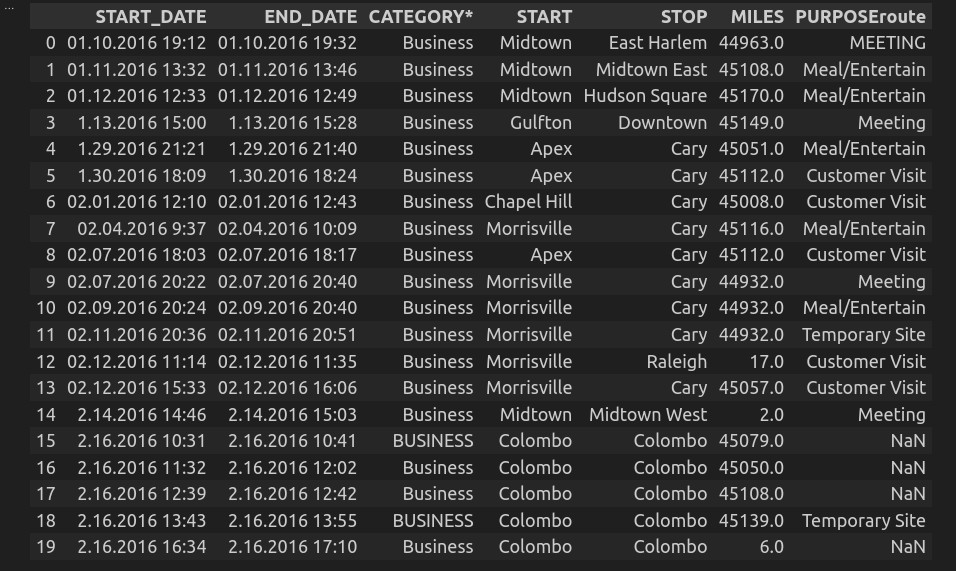


Рисунок 1 - Вывод первых 20 элементов

Столбцы:

1. START\_DATE и END\_DATE: Эти столбцы содержат дату и время начала и завершения поездки. Они указывают, когда началась и завершилась каждая поездка.
2. CATEGORY: Этот столбец обозначает категорию поездки. В данном случае, больше поездок "Business", что указывает на то, что она связана с деловой деятельностью.
3. START и STOP: Эти столбцы указывают начальное и конечное местоположение поездки. Они указывают, откуда и куда направлялась поездка.
4. MILES: Этот столбец содержит информацию о количестве миль, пройденных во время поездки. Это может быть полезным для расчета расстояний и затрат на топливо.
5. PURPOSE: Этот столбец описывает цель поездки. Он указывает, для чего была совершена поездка, например, "MEETING" (встреча), "Meal/Entertain" (питание/развлечение), "Customer Visit" (визит к клиенту) и т. д.

Предметная область этой таблицы связана с учетом и анализом служебных поездок. Скорее всего, она используется для отслеживания затрат на бизнес-поездки, анализа расходов на топливо, определения целей и местоположение поездок, а также для управления служебными маршрутами и встречами.

Далее с помощью команды df.info() была выведена информацию о DataFrame (Рисунок 2).

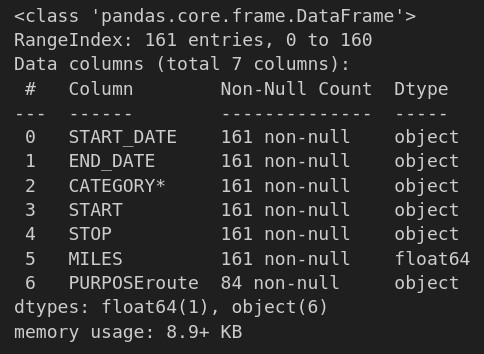


Рисунок 2 - Информация о DataFrame

Результат выполнения метода info() показывает следующую информацию о DataFrame:

1. Всего строк: 161.
2. Количество непустых (non-null) значений для каждого столбца.
3. Типы данных для каждого столбца.

Из этой информации видно следующее:

* Столбцы "START\_DATE", "END\_DATE", "CATEGORY\*", "START", "STOP" и

"PURPOSEroute" содержат объекты (строки).

- Столбец "MILES" содержит числовые значения с типом данных float64.

- Столбец "PURPOSEroute" имеет некоторые пропущенные значения, так как количество непустых значений меньше общего числа строк.

С помощью команды df.describe() были оценены числовые столбцы (Рисунок 3).

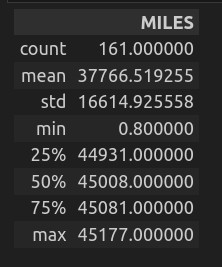


Рисунок 3 - Числовые столбцы

Результат выполнения метода describe() для числового столбца "MILES" выглядит следующим образом:

- count: Количество непустых (non-null) значений в столбце - 161.

- mean: Среднее значение - 37766.519255.

- std: Стандартное отклонение - 16614.925558.

- min: Минимальное значение - 0.800000.

- 25%: 25-й процентиль - 44931.000000.

- 50%: Медианное значение (50-й процентиль) - 45008.000000.

- 75%: 75-й процентиль - 45081.000000.

- max: Максимальное значение - 45177.000000.

Эти статистические показатели позволяют оценить основные характеристики числового столбца "MILES", такие как среднее значение, разброс данных (стандартное отклонение), минимальное и максимальное значения, а также квантили, что может быть полезно при анализе данных и поиске выбросов.

Далее были выведены на экран названия столбцов с помощью df.columns (Рисунок 4).



Рисунок 4 - Вывод названия столбцов

Проблема заключается в том, что названия не имеют единого формата. Для этого были переименованы столбцы. Код продемонстрирован в листинге 2, результат — на рисунке 5.

Листинг 2 - Код замены названий столбцов

df.columns = ['START\_DATE', 'END\_DATE', 'CATEGORY', 'START', 'STOP', 'MILES', 'PURPOSE\_ROUTE']

df.columns

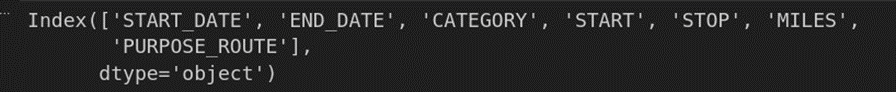


Рисунок 5 - Изменения названия столбцов

С помощью метода fillna() были заменены пропуски на 'Unknown' (Листинг 3).

Листинг 3 - Код замены пропусков

df['PURPOSE\_ROUTE'].fillna('Unknown', inplace=True)

Результат работы данного кода продемонстрирован на рисунке 6.

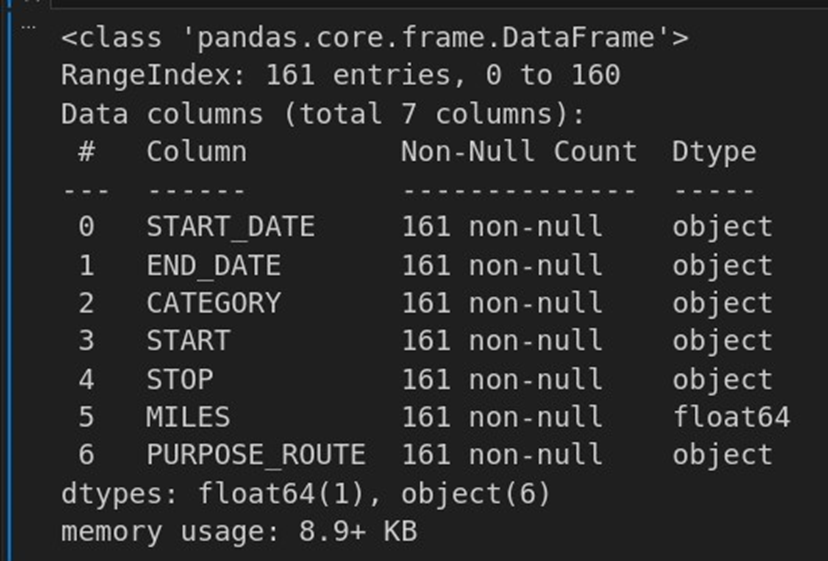


Рисунок 6 - Замена пропусков

Далее была проведена проверка на явные дубликаты с помощью метода df[df.duplicated()] (Рисунок 7).

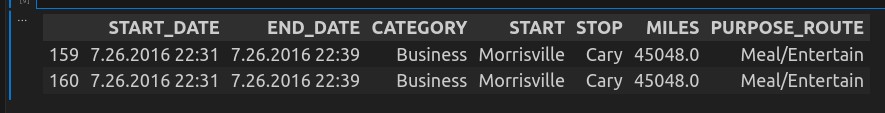


Рисунок 7 - Проверка на дубликаты

На рисунке 7 видно, что есть еще две строки идентичные другой, с помощью drop\_duplicates(inplace=True) избавились от них, после была проведена проверка на дубликаты еще раз. На рисунке 8 видно, что больше дубликатов не было обнаружено.

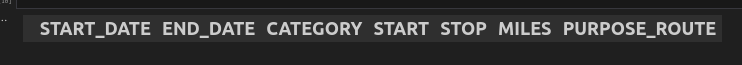


Рисунок 8 - Результат удаление дубликатов

Была проведена проверка на неявные дубликаты (различные написания одного и того же) при помощи функции df['CATEGORY'].unique() для столбцов с текстовым типом данных (Рисунок 9). Для начала проверен столбец CATEGORY.

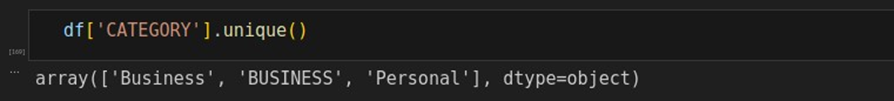


Рисунок 9 - Просмотр на наличие уникальных значение

На рисунке 9 видим, что Business пишется в двух вариациях. С помощью replace() значения были изменены на общий вид (Листинг 4).

df['CATEGORY'] = df['CATEGORY'].replace('BUSINESS',

'Business')

Листинг 4 - Код замены значений.

Результат работы продемонстрирован на рисунке 10.



Рисунок 10 - Замена неявных дубликатов

Тоже самое было проделано и для столбца PURPOSE\_ROUTE, другие не рассматривались, так как это или дата, или название (Рисунок 11). Проверка на неявные дубликаты ограничивается текстовыми данными из-за синтаксических различий, чувствительности к регистру и сложностей с точностью числовых данных. Для текстовых данных проверка на дубликаты более надежна и предсказуема.

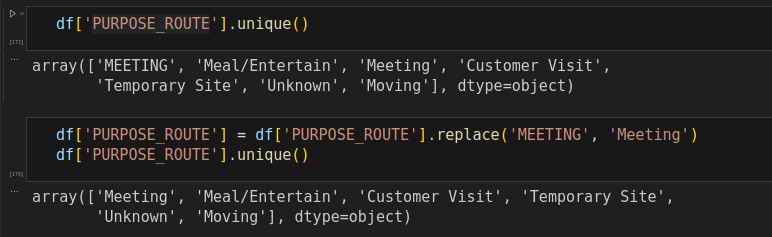


Рисунок 11 - Замена неявных дубликатов для другого столбца

У столбцов с датой был изменен тип данных на datetime (Листинг 5). Листинг 5 - Код изменения типа данных

df['START\_DATE'] = pandas.to\_datetime(df['START\_DATE'], format='%m.%d.%Y %H:%M')

df['END\_DATE'] = pandas.to\_datetime(df['END\_DATE'], format='%m.%d.%Y %H:%M')

df.dtypes

Результат работы продемонстрирован на рисунке 12.

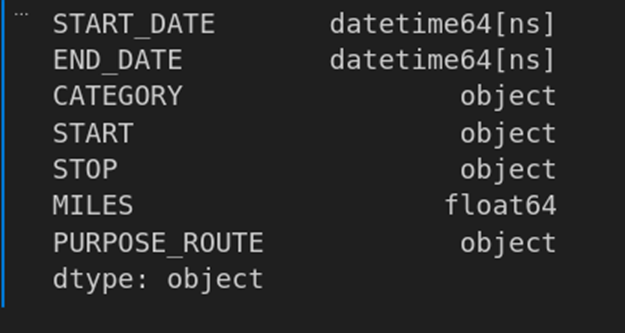


Рисунок 12 - Изменение типов данных

Далее были выполнены индивидуальные задания.

Задание 1

Группировка - CATEGORY и количество поездок каждого типа (по цели маршрута). Код продемонстрирован в листинге 6.

Листинг 6 - Код 1 задания

df.groupby(['CATEGORY', 'PURPOSE\_ROUTE']) ['PURPOSE\_ROUTE'].count()

Данный код подсчитывает с какой целью часто совершают поездки, исходя из категории. Видно, что в категории деловых поездок, наиболее популярная цель поездки - обед, а в категории личных поездок - встреча (не считая пустых ячеек). Результат работы продемонстрирован на рисунке 13.

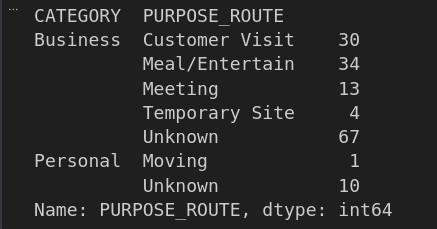


Рисунок 13 - Результат задания 1

Этот анализ помогает понять, какие категории и цели маршрутов наиболее распространены или часто встречаются в данных, что может быть полезной информацией для принятия бизнес-решений или дальнейшего анализа.

Задание 2

Группировка - CATEGORY и количество поездок каждого типа (по цели маршрута). Создать датафрейм. Переименовать столбец с количеством в “сount”. Отсортировать по убыванию столбца “count”.

Код продемонстрирован в листинге 7.

Листинг 7 - Код задания 2

df1 = df.groupby(['CATEGORY', 'PURPOSE\_ROUTE'])

['PURPOSE\_ROUTE'].count().reset\_index(name

='count').sort\_values('count', ascending=False) df1

Была создана группировка как в предыдущем задании, столбец с полученным значением назван – count, также была применена функция sort\_values для данного столбца по убыванию. Результат работы продемонстрирован на рисунке 14.

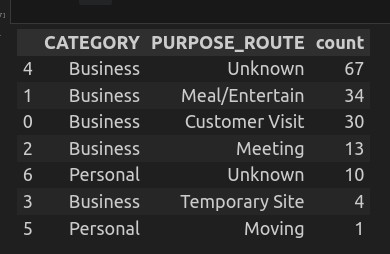


Рисунок 14 - Результат задания 2

Из выведенных данных можно заметить, что очень часто встречается неизвестная цель поездки, для более грамотного учета данных, следует обязать заполнять эти данные.

Анализ таких данных может помочь определить, какие цели маршрута наиболее популярны или наименее популярны в каждой из категорий поездок. Например, эти данные могут быть полезны для бизнес-аналитики при определении, на какие виды мероприятий или услуг следует сосредотачивать усилия или ресурсы.

Задание 3

Сводная таблица (pivot\_table) - средняя количество пройденных миль по каждой цели поездки (PURPOSEroute). Отсортировать по убыванию столбца MILES. Округлить значение до двух знаков.

Код продемонстрирован в листинге 8.

Листинг 8 - Код задания 3

df\_pivot = pandas.pivot\_table(df, values='MILES', index='PURPOSE\_ROUTE', aggfunc='mean')

df\_pivot['MILES'] = df\_pivot['MILES'].round(2) df\_pivot.sort\_values(by='MILES', ascending=False,

inplace=True) df\_pivot

Для начала была создана сводная таблиця, используя - pivot\_table. Далее столбец MILES округлен до двух знаков после запятой и отсортирован по убыванию.

Результат работы продемонстрирован на рисунке 15.

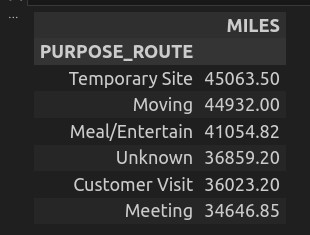


Рисунок 15 - Результат задания 3

Самая большая средняя длина поездки (около 45063.50 миль) наблюдается для цели маршрута "Temporary Site". Это может указывать на длительные поездки, связанные с временными рабочими объектами или местоположениями.

Анализ средней длины поездок по целям маршрута может быть полезным при определении того, какие типы поездок являются наиболее долгими или короткими, что может иметь значение для планирования и оценки ресурсов или стоимости поездок.

Задание 4

Сводная таблица (pivot\_table) - средняя количество пройденных миль по каждой цели поездки (PURPOSEroute) - столбцы и каждой категории - строки. Отсортировать по убыванию столбца CATEGORY.Код продемонстрирован в листинге 9.

df\_pivot2 = pandas.pivot\_table(df, columns='PURPOSE\_ROUTE', values='MILES', index='CATEGORY', aggfunc='mean')

df\_pivot2.sort\_values(by='CATEGORY', ascending=False) df\_pivot2

Листинг 9 - Код задания 4

Проделана таже работа, что и в прошлом задании, но теперь в pivot\_table передается массив столбцов. Сортировка же для символьных типов данных идет по ASCII таблице (по алфавитному порядку). Результат работы продемонстрирован на рисунке 16.

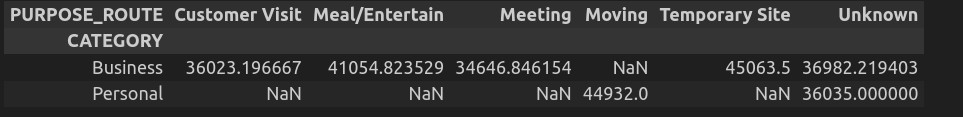


Рисунок 16 - Результат задания 4

Дополнительное задание

Добавлен столбец с длительностью поездки, выбран только бизнес-класс. Составлена сводная таблица с целью маршрута и его длительностью. Код продемонстрирован в листинге 10.

df\_pivot3 = df[df["CATEGORY"] == "Business"].copy() df\_pivot3["TIME\_IN\_ROAD"] = (df\_pivot3["END\_DATE"] -

df\_pivot3["START\_DATE"]).dt.total\_seconds() / 60 df\_pivot3 = pandas.pivot\_table(df\_pivot3,

values='TIME\_IN\_ROAD', index='PURPOSE\_ROUTE', aggfunc='mean')

df\_pivot3["TIME\_IN\_ROAD"] = df\_pivot3["TIME\_IN\_ROAD"].round(2)

df\_pivot3

Листинг 10 - Код дополнительного задания

Результат работы продемонстрирован на рисунке 17.

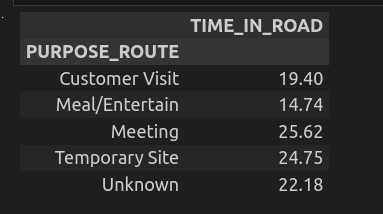


Рисунок 16 - Результат задания 4

Самое короткое среднее время в пути наблюдается для цели маршрута "Meal/Entertain" (14.74 минут). Это может означать, что поездки, связанные с обедами или развлечениями, обычно не требуют много времени на перемещение.

Анализ таких данных может быть полезным при планировании или оценке времени, затрачиваемого на деловые поездки, и может помочь определить, какие типы поездок часто требуют большего времени в пути.

# **Вывод**

В ходе анализа конкретного набора данных о поездках была проведена детальная работа, направленная на выявление основных тенденций и характеристик маршрутов в зависимости от их целей и категорий. В результате анализа были сделаны следующие выводы:

1. Цели поездок и их популярность: из проведенного анализа видно, что наиболее распространенными целями деловых поездок являются обеды, тогда как в случае личных поездок — это встречи. Однако также обнаружены пустые ячейки, что указывает на отсутствие информации о целях некоторых поездок. Эта информация может быть ценной для компании, так как она может помочь оптимизировать услуги в соответствии с популярными потребностями клиентов.
2. Наиболее часто встречающиеся типы поездок: Неизвестные цели поездок были выделены как наиболее часто встречающиеся. Это выявление указывает на необходимость более точного сбора данных о целях поездок. Подобная информация может помочь в улучшении алгоритмов или методов сбора данных.
3. Средняя длина поездок в зависимости от целей: По результатам анализа стало ясно, что цель поездки "Temporary Site" обычно связана с длительными поездками, возможно, связанными с работой на временных объектах или местоположениях. Эта информация может быть важной для бизнеса, который занимается планированием и оптимизацией маршрутов.

В целом, проведенный анализ данных позволяет сделать выводы, которые могут быть весьма полезны для бизнеса в плане оптимизации маршрутов, понимания потребностей клиентов и улучшения качества предоставляемых услуг. Однако, также важно отметить, что точность данных играет ключевую роль в получении более надежных результатов анализа, и поэтому рекомендуется уделить внимание сбору данных для будущих исследований и анализа.