|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ (ИУ5)\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

по дисциплине \_\_\_\_\_\_\_\_\_\_\_\_\_ Оперативный анализ данных \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

по теме\_\_\_\_\_\_ Анализ данных по базе данных «Phishing URL detection»\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент ИУ5-52Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** ЧугуновА.В.

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** МаслениковК.Ю.

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Маслеников К.Ю.

(Подпись, дата) (И.О.Фамилия)

*2024 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_Анализ данных по базе данных «Phishing URL detection» \_\_

Студент группы \_\_\_\_\_\_\_ИУ5-52Б\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Чугунов Александр Владимирович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Исследовательская\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики(кафедра,предприятие,НИР) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_НИР\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к 3 нед., 50% к 9 нед., 75% к 12 нед., 100% к 15 нед.

***Техническое задание***  Спроектировать систему анализа и визуализации по теме « HR Dataset», провести визуализацию полученных данных посредством языка программирования Python

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на 23 листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «14» сентября 2024 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Маслеников К.Ю.

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Чугунов А.В.

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**Аннотация**

По заданию требуется проанализировать базу данных, очистить её от

ненужных данных, также нужно выдвинуть гипотезы, проанализировать их и составить графики по данным гипотезам. Для данного задания была выбрана база данных «Phishing URL detection». В ходе работы будут использоваться такие библиотеки как: «Pandas» и «Matplotlib». Анализ будет проводиться через программу «Jupyter Notebook».

Также будет производиться корреляционный анализ, агрегирование данных, оптимизация памяти, удаление дубликатов, очистка данных. Будет осуществлён анализ гипотез по известным данным и построены соответствующие графики и схемы.

В работе будут рассмотрены данные о фишинговых url, которые могут быть использованы для анализа факторов, вляиющих на вероятность использование url адреса для фишинга. В конце, будет сделан вывод о проделанной работе с данным датасетом.

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 4](#__RefHeading___Toc1127_1261885667)

[Определение данных для анализа 5](#__RefHeading___Toc1129_1261885667)

[1. Описание данных 7](#__RefHeading___Toc1131_1261885667)

[2. Формулирование гипотез 7](#__RefHeading___Toc1133_1261885667)

[3. Подготовка данных для работы 9](#__RefHeading___Toc1135_1261885667)

[4. Изучение общей информации 10](#__RefHeading___Toc1137_1261885667)

[5. Исследовательский анализ данных 12](#__RefHeading___Toc1139_1261885667)

[6. Анализ корреляционной матрицы 22](#__RefHeading___Toc1141_1261885667)

[ЗАКЛЮЧЕНИЕ 24](#__RefHeading___Toc1143_1261885667)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 25](#__RefHeading___Toc1145_1261885667)

# **ВВЕДЕНИЕ**

Необходимо исследовать данные и произвести анализ факторов, влияющих на классификацию URL-адресов как фишинговых или легитимных.

Цели:

1. определение данных;
2. формулирование гипотез;
3. загрузка данных в Python;
4. проверка данных;
5. очистка данных;
6. преобразование данных;
7. выбор данных для анализа;
8. агрегирование данных;
9. визуализация данных;
10. подтверждение или опровержение поставленных гипотез;
11. формулирование ограничений и выводов.

# **Определение данных для анализа**

В качестве данных для анализа был выбран датасет «Phishing URL Detection».  
 Этот датасет содержит информацию о различных характеристиках URL-адресов, которые могут быть использованы для анализа факторов, влияющих на классификацию URL как фишинговых или легитимных. Датасет включает такие признаки, как длина URL, количество точек, наличие количество цифр в URL, количество субдоменов, использование HTTPS, количество специальных символов, случайность домена и другие характеристики. Кроме того, в датасете указано, является ли URL фишинговым или легитимным.

# **Описание данных**

Для анализа были собраны данные о различных характеристиках URL-адресов, включенных в датасет. В наборе данных содержатся следующие поля:

* url\_length - длина URL.
* hostname\_length - длина хоста.
* total\_of. - количество точек в URL.
* total\_of- - количество дефисов в URL.
* total\_of@ - количество символов "@" в URL (может быть признаком фишинга).
* total\_of? - количество знаков вопроса в URL.
* https\_token - наличие HTTPS в URL.
* ratio\_digits\_url - соотношение цифр в URL.
* ratio\_digits\_host - соотношение цифр в хосте.
* abnormal\_subdomain - наличие аномальных поддоменов.
* nb\_subdomains - количество поддоменов.
* random\_domain - случайность домена (может быть индикатором фишинга).
* shortening\_service - использование сервисов сокращения ссылок.
* nb\_redirection - количество редиректов (может указывать на фишинг).
* phish\_hints - признаки фишинга.
* **status —** фишинг / легитимный

# **Формулирование гипотез**

Гипотеза 1: URL-адреса с большей длиной имеют более высокую вероятность быть фишинговыми.

Гипотеза 2: URL-адреса, содержащие IP-адрес вместо доменного имени, чаще классифицируются как фишинговые.

Гипотеза 3: Наличие большого количества специальных символов в URL коррелирует с более высокой вероятностью фишинга.

Гипотеза 4: URL-адреса, использующие HTTPS, реже классифицируются как фишинговые.

Гипотеза 5: URL-адреса из менее популярных доменных зон (TLD), чаще оказываются фишинговыми по сравнению с популярными доменными зонами.

# **Подготовка данных для работы**

Загружаем датасет и подключаем необходимые библиотеки.  
Выбираем из датасета поля, необходимые для анализа.

Проводим очистку данных.

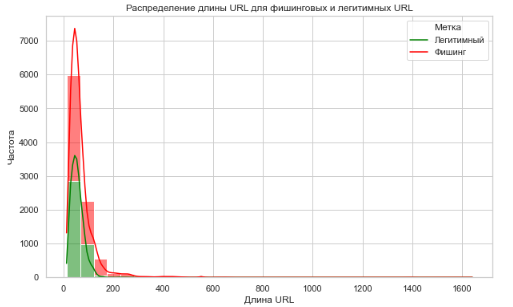


# **Изучение общей информации**

В данном разделе представлен общий анализ данных о фишинговых URL. Также представлена визуализация распределения длины URL относительно частоты их появления в датасете, а также распределение меток, показывающих, являются ли URL фишинговыми или легитимными.

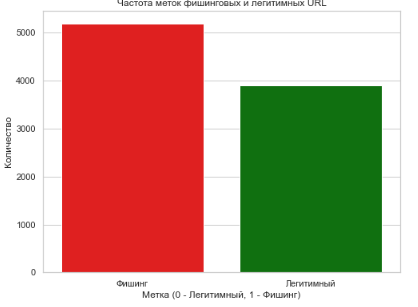


Первая гистограмма позволяет наблюдать, как длина URL распределяется среди фишинговых и легитимных сайтов.



**Вывод:** Гистограмма показывает, что длина фишинговых url обычно несколько больше длины легитимных url.

Вторая гистограмма показывает частоту меток фишинговых и легитимных url.

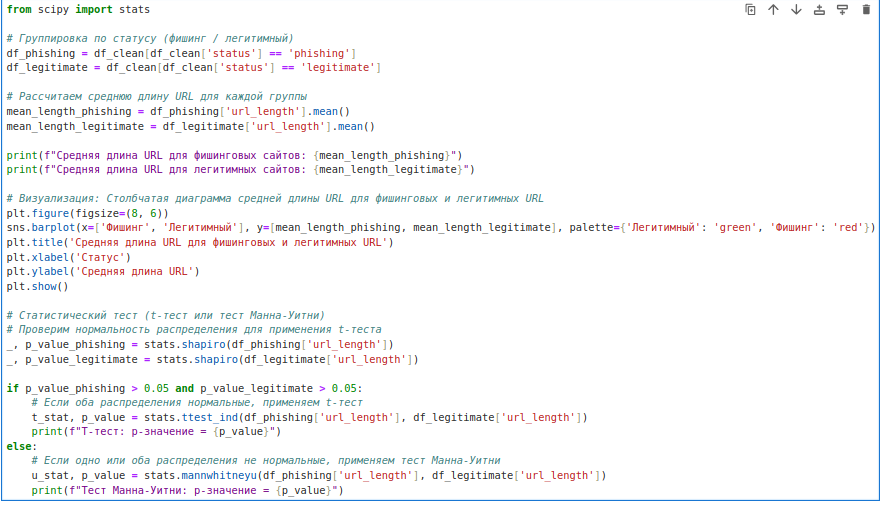


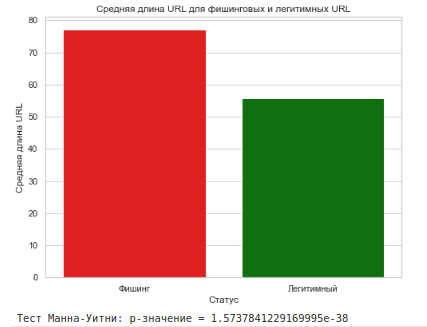
**Вывод**: Гистограмма показывает, что в выбранном датасете фишинговые url преобладают над легитимными.

# **Исследовательский анализ данных**

**Гипотеза 1: URL-адреса с большей длиной имеют более высокую вероятность быть фишинговыми.**

Для проверки гипотезы можно использовать тест Манна-Уитни и диаграмму barplot, чтобы увидеть, есть ли связь между длиной url и его легитимностью

Визуализируем полученные данные с помощью диаграммы:



**Вывод**:

Диаграмма показывает, что фишинговые url в среднем имеют большую длину: 77 символов в фишинговых url против 55 символов в легитимных url.

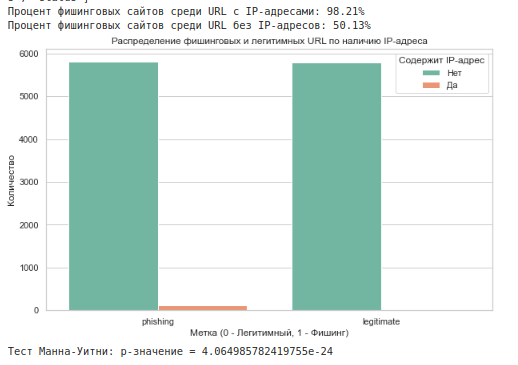
При проведении теста на статистическую значимость разницы выборок также был получен результат, свидетельствующий о том, что существует статистически значимое различие в длине URL между фишинговыми и легитимными сайтами.

Таким образом, гипотеза о том, что URL-адреса с большей длиной имеют более высокую вероятность быть фишинговыми, **подтверждается**.

**Гипотеза 2: URL-адреса, содержащие IP-адрес вместо доменного имени, чаще классифицируются как фишинговые.**

Для анализа гипотезы, визуализируем данные с помощью столбчатой диаграммы. Для определения является ли url ip-адресом используем регулярное выражение.





**Вывод**:

Процент фишинговых сайтов среди URL с IP-адресами составил 98.21%, что говорит о крайне высокой вероятности того, что сайт с url в виде ip-адреса является фишинговым.

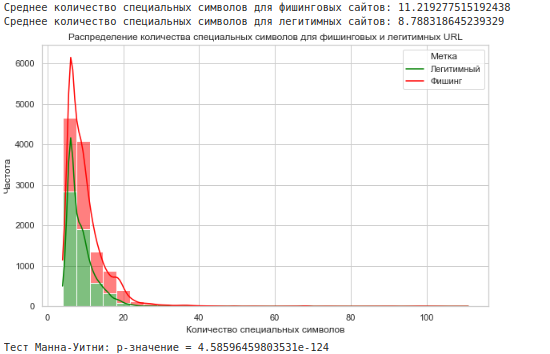
Тест Манна-Уитни также показал крайне малое значение, что говорит о том, что статистически сайты с url-адресами в виде ip с большей вероятностью являются фишинговыми.

Таким образом, гипотеза о том, что URL-адреса, содержащие IP-адрес вместо доменного имени, чаще классифицируются как фишинговые, **подтверждается.**

**Гипотеза 3: Наличие большого количества специальных символов в URL коррелирует с более высокой вероятностью фишинга.**

Для проверки гипотезы выделим количество специальных символов, разделим на группу и визуализируем для сравнения двух групп.



****

**Вывод:**

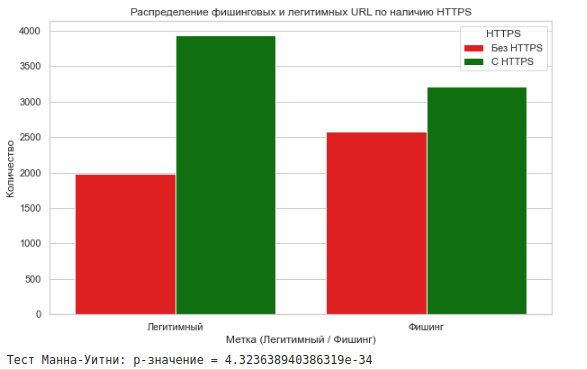
По графику видно, что для фишинговых url характерно несколько большее количество специальных символов, чем для легитимных.

Тест Манна-Уитни показал крайне малое значение, что также свидетельствует о том, что существует статистическая разница в количестве специальных символов между фишинговыми и легитимными url.

Таким образом, гипотеза о том, что наличие большого количества специальных символов в URL коррелирует с более высокой вероятностью фишинга, **подтверждается.**

**Гипотеза 4: URL-адреса, использующие HTTPS, реже классифицируются как фишинговые.**





**Вывод**:

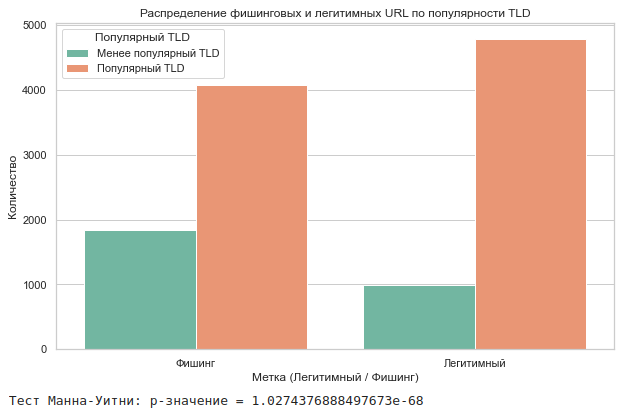
По гистограмме видно, что доля сайтов, использующих https, среди легитимных относительно сайтов без https больше, чем такая же доля фишинговых сайтов.

Результаты теста Манна-Уитни показали крайне малое p-значение, что указывает на статистически значимую разницу между фишинговыми и легитимными URL по наличию HTTPS. Это означает, что гипотеза о том, что URL-адреса, использующие HTTPS, реже классифицируются как фишинговые, **подтверждается**.

**Гипотеза 5: URL-адреса из менее популярных доменных зон (TLD), чаще оказываются фишинговыми по сравнению с популярными доменными зонами.**

Сгруппируем url по доменной зоне, по статусу и проведем статистический тест.





**Вывод**:

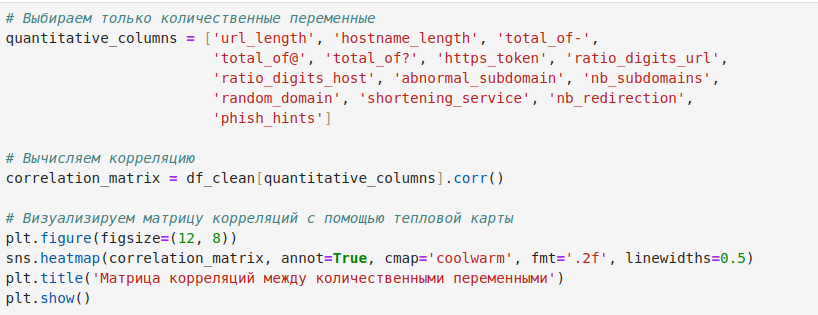
По гистограмме видно, что доля фишинговых сайтов, использующих менее популярные TLD относительно популярных, больше той же доли легитимных сайтов.

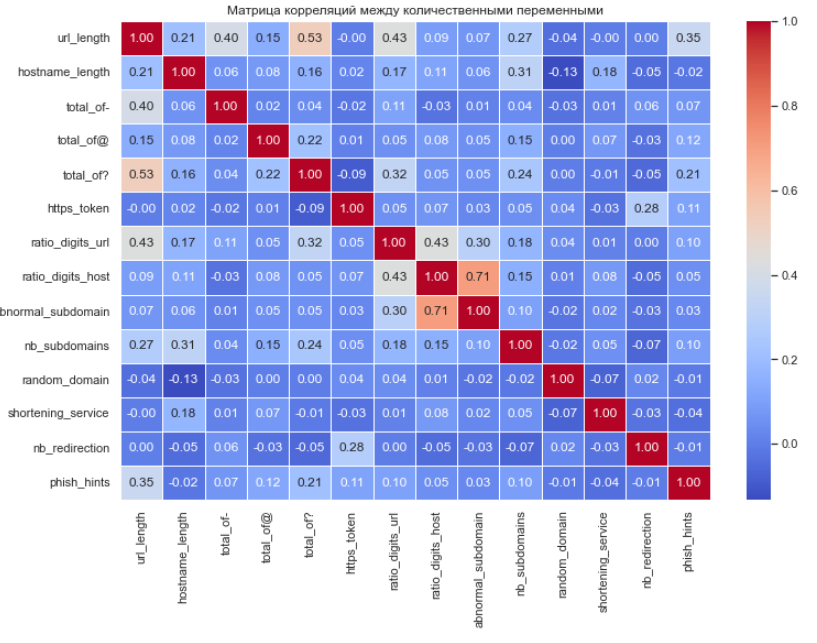
Тест Манна-Уитни показал крайне малое значение, что указывает на статистически значимую разницу между TLD фишинговых и легитимных url.

Таким образом, гипотеза о том, что URL-адреса из менее популярных доменных зон (TLD), чаще оказываются фишинговыми по сравнению с популярными доменными зонами, **подтверждается.**

# **Анализ корреляционной матрицы**

Построим матрицу корреляций на основе количественных переменных датасета.





**Вывод**:

На основе построенной матрицы корреляций можно выделить несколько важных наблюдений.

Прежде всего, следует отметить, что большая часть количественных переменных в датасете не имеет сильной корреляции друг с другом. Например, такие переменные, как количество точек (total\_of.), дефисов (total\_of-), символов "@" (total\_of@), вопросительных знаков (total\_of?) в URL не коррелируют с другими признаками в значимой степени. Это может свидетельствовать о том, что различные специальные символы в URL могут быть полезными индикаторами для фишинга, но не всегда взаимосвязаны между собой.

Тем не менее, стоит выделить важную корреляцию, которая наблюдается количеством цифр URL (url\_length) и переменной abnormal\_subdomain. Коэффициент корреляции между этими двумя переменными составляет 0.71, что указывает на сильную положительную связь. Вероятно, в датасете количество цифр в субдомене является критерием его аномальности.

Кроме того, можно отметить, что переменные, такие как соотношение цифр в URL (ratio\_digits\_url) и в хосте (ratio\_digits\_host), имеют некоторые корреляции с длиной URL. Это логично, так как URL с большим количеством символов, в том числе цифр и других знаков, как правило, будут длиннее. Однако корреляции здесь не настолько сильные, чтобы можно было сделать однозначные выводы.

В целом, большинство переменных в датасете показывают низкую корреляцию друг с другом. Основным фактором, влияющим на корреляции, является длина URL, что также подтверждается высокими коэффициентами корреляции с рядом других переменных. Это может означать, что длина URL — один из ключевых факторов при анализе фишинговых ссылок.

# **ЗАКЛЮЧЕНИЕ**

В ходе научно-исследовательской работы был проведен анализ базы данных «Phishing URL detection», выявлены важные статистические данные, которые могут стать мощным инструментом для изучения различных аспектов управления сотрудниками компании и удержания кадров.

В ходе работы были освоены инструменты анализа данных, включая библиотеки Pandas, Seaborn и Matplotlib в среде программирования Python (Jupyter Notebook). В результате были получены новые навыки в области анализа данных и построения визуализаций.

Была осуществлена оптимизация использования памяти, агрегирование данных, визуализация данных с помощью графиков и диаграмм, а также были проведены тесты для определения статистической значимости данных, что в дальнейшем позволило провести необходимые в ходе работы исследования и получить знания в предметной области.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Методические указания по программному обеспечению «Pandas»;
2. Методические указания по программному обеспечению «Seaborn»;
3. Методические указания по программному обеспечению «Plotly»;
4. «Pandas. Работа с данными» (2020), Автор: Абдрахманов М. И.;
5. «Python. Визуализация данных: Matplotlib, Seaborn, Mayavi»;
6. [https://devpractice.ru](https://devpractice.ru/);
7. [https://coderoad.wiki](https://coderoad.wiki/);
8. [https://www.delftstack.com](https://www.delftstack.com/).