РК 1 Зоров Владислав ИУ5-22М, вариант 4, задачи 4, 24

Для произвольной колонки данных построить гистограмму.

Задача 4 Для набора данных проведите кодирование одного (произвольного) категориального признака с использованием метода "label encoding".

Задача 24 Для набора данных для одного (произвольного) числового признака проведите обнаружение и удаление выбросов на основе 5% и 95% квантилей.

Задача 4

In [20]:

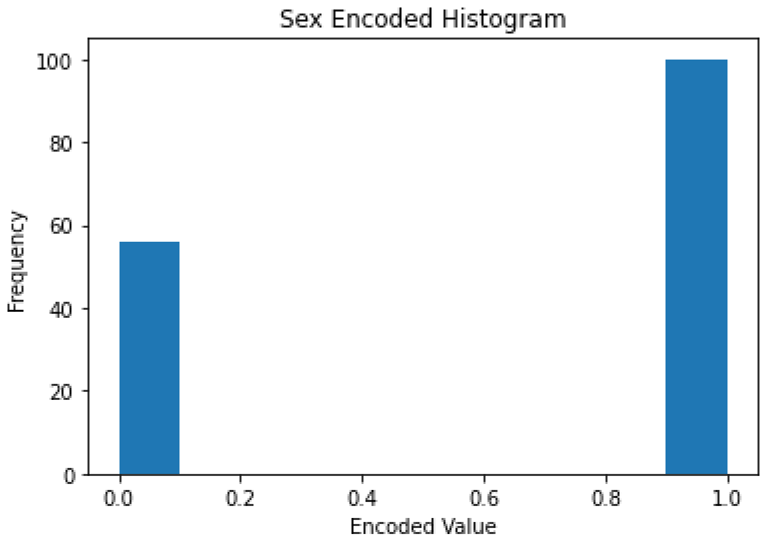
```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder

# Загрузка данных
data = pd.read_csv('E:/titanic.csv')
#Для кодирования категориального признака воспользуюсь методом "label encoding", кодирую 

print(data)
# Создание объекта LabelEncoder
le = LabelEncoder()

# Кодирование категориального признака
data['sex_encoded'] = le.fit_transform(data['Sex'])

print(data)
# Был закодирован признак Sex, построю гистограмму для него
# Наличие пропуска кают у некоторых пассажиров не влияет на результат кодировки возраста
```

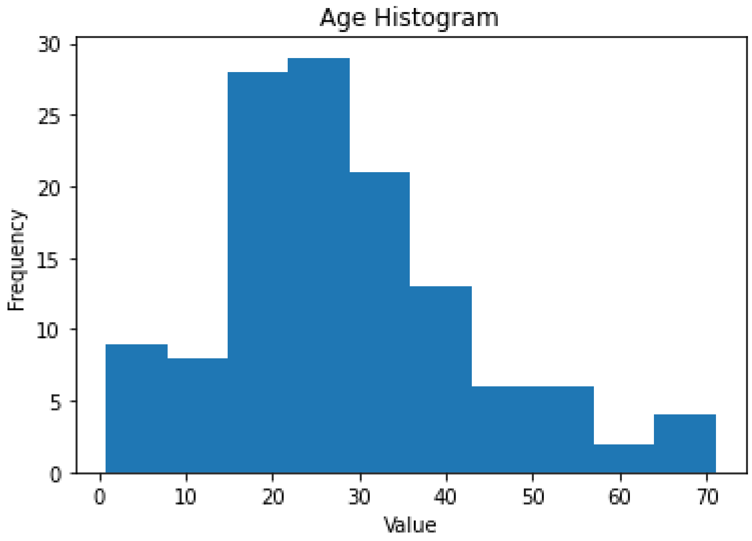
```
     PassengerId  Survived  Pclass         Lname  \
0              1         0       3        Braund   
1              2         1       1       Cumings   
2              3         1       3     Heikkinen   
3              4         1       1      Futrelle   
4              5         0       3         Allen   
..           ...       ...     ...           ...   
151          152         1       1         Pears   
152          153         0       3           Meo   
153          154         0       3  van Billiard   
154          155         0       3         Olsen   
155          156         0       1      Williams   

                                           Name     Sex   Age  SibSp  Parch  \
0                               Mr. Owen Harris    male  22.0      1      0   
1    Mrs. John Bradley (Florence Briggs Thayer)  female  38.0      1      0   
2                                   Miss. Laina  female  26.0      0      0   
3            Mrs. Jacques Heath (Lily May Peel)  female  35.0      1      0   
4                             Mr. William Henry    male  35.0      0      0   
..                                          ...     ...   ...    ...    ...   
151                 Mrs. Thomas (Edith Wearne)  female  22.0      1      0   
152                                 Mr. Alfonzo    male  55.5      0      0   
153                           Mr. Austin Blyler    male  40.5      0      2   
154                              Mr. Ole Martin    male   NaN      0      0   
155                           Mr. Charles Duane    male  51.0      0      1   
```

```
                Ticket     Fare Cabin Embarked
0            A/5 21171   7.2500   NaN        S
1             PC 17599  71.2833   C85        C
2     STON/O2. 3101282   7.9250   NaN        S
3               113803  53.1000  C123        S
4               373450   8.0500   NaN        S
..                 ...      ...   ...      ...
151             113776  66.6000    C2        S
152         A.5. 11206   8.0500   NaN        S
153           A/5. 851  14.5000   NaN        S
154          Fa 265302   7.3125   NaN        S
155           PC 17597  61.3792   NaN        C

[156 rows x 13 columns]
     PassengerId  Survived  Pclass         Lname  \
0              1         0       3        Braund
1              2         1       1       Cumings
2              3         1       3     Heikkinen
3              4         1       1      Futrelle
4              5         0       3         Allen
..           ...       ...     ...           ...
151          152         1       1         Pears
152          153         0       3           Meo
153          154         0       3  van Billiard
154          155         0       3         Olsen
155          156         0       1      Williams

                                           Name     Sex   Age  SibSp  Parch  \
0                              Mr. Owen Harris    male  22.0      1      0
1     Mrs. John Bradley (Florence Briggs Thayer)  female  38.0      1      0
2                                   Miss. Laina  female  26.0      0      0
3             Mrs. Jacques Heath (Lily May Peel)  female  35.0      1      0
4                             Mr. William Henry    male  35.0      0      0
..                                          ...     ...   ...    ...    ...
151                   Mrs. Thomas (Edith Wearne)  female  22.0      1      0
152                                 Mr. Alfonzo    male  55.5      0      0
153                           Mr. Austin Blyler    male  40.5      0      2
154                              Mr. Ole Martin    male   NaN      0      0
155                          Mr. Charles Duane    male  51.0      0      1

                Ticket     Fare Cabin Embarked  sex_encoded
0            A/5 21171   7.2500   NaN        S            1
1             PC 17599  71.2833   C85        C            0
2     STON/O2. 3101282   7.9250   NaN        S            0
3               113803  53.1000  C123        S            0
4               373450   8.0500   NaN        S            1
..                 ...      ...   ...      ...          ...
151             113776  66.6000    C2        S            0
152         A.5. 11206   8.0500   NaN        S            1
153           A/5. 851  14.5000   NaN        S            1
154          Fa 265302   7.3125   NaN        S            1
155           PC 17597  61.3792   NaN        C            1

[156 rows x 14 columns]
```

In [22]:

```
# Построение гистограммы
plt.hist(data['sex_encoded'])
plt.title('Sex Encoded Histogram')
plt.xlabel('Encoded Value')
plt.ylabel('Frequency')
plt.show()

plt.hist(data['Age'])
plt.title('Age Histogram')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.show()
```





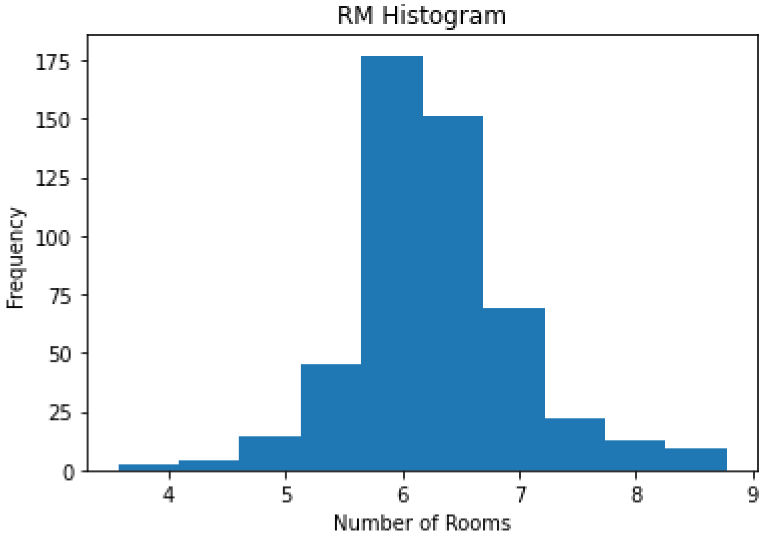
Задача 24

In [23]:

```
import pandas as pd
import matplotlib.pyplot as plt

# Загрузка данных
data = pd.read_csv('E:/BostonHousing.csv')
print(data)




plt.hist(data['rm'])
plt.title('RM Histogram')
plt.xlabel('Number of Rooms')
plt.ylabel('Frequency')
plt.show()
```

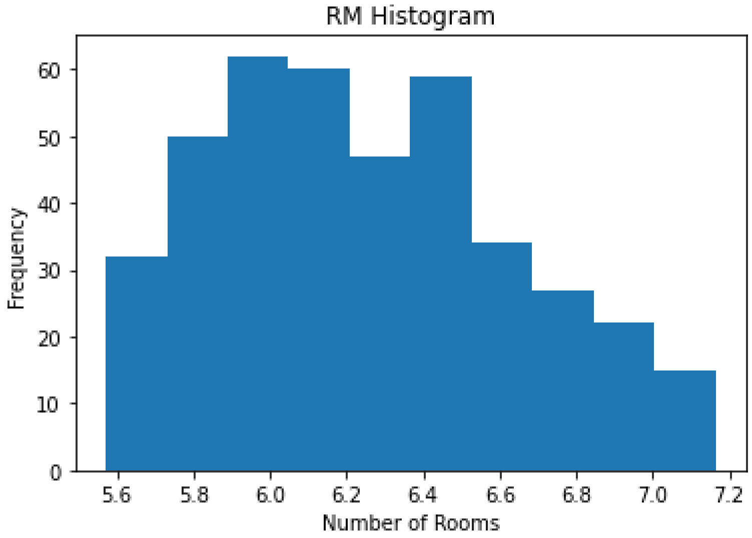
```
        crim    zn  indus  chas    nox     rm   age     dis  rad  tax  \
0    0.00632  18.0   2.31     0  0.538  6.575  65.2  4.0900    1  296
1    0.02731   0.0   7.07     0  0.469  6.421  78.9  4.9671    2  242
2    0.02729   0.0   7.07     0  0.469  7.185  61.1  4.9671    2  242
3    0.03237   0.0   2.18     0  0.458  6.998  45.8  6.0622    3  222
4    0.06905   0.0   2.18     0  0.458  7.147  54.2  6.0622    3  222
```

```
..       ...  ...    ...  ...    ...    ...   ...     ...  ...  ...
501  0.06263  0.0  11.93    0  0.573  6.593  69.1  2.4786    1  273
502  0.04527  0.0  11.93    0  0.573  6.120  76.7  2.2875    1  273
503  0.06076  0.0  11.93    0  0.573  6.976  91.0  2.1675    1  273
504  0.10959  0.0  11.93    0  0.573  6.794  89.3  2.3889    1  273
505  0.04741  0.0  11.93    0  0.573  6.030  80.8  2.5050    1  273

     ptratio       b  lstat  medv
0       15.3  396.90   4.98  24.0
1       17.8  396.90   9.14  21.6
2       17.8  392.83   4.03  34.7
3       18.7  394.63   2.94  33.4
4       18.7  396.90   5.33  36.2
..       ...     ...    ...   ...
501     21.0  391.99   9.67  22.4
502     21.0  396.90   9.08  20.6
503     21.0  396.90   5.64  23.9
504     21.0  393.45   6.48  22.0
505     21.0  396.90   7.88  11.9

[506 rows x 14 columns]
```



In [18]:

```python
# Определение 5% и 95% квантилей
q05 = data['rm'].quantile(0.05)
q95 = data['rm'].quantile(0.95)


data = data[(data['rm'] >= q05) & (data['rm'] <= q95)]



# Построение гистограммы
plt.hist(data['rm'])
plt.title('RM Histogram')
plt.xlabel('Number of Rooms')
plt.ylabel('Frequency')
plt.show()
```



In [24]:

```
!pip install Pyppeteer
```

```
Collecting Pyppeteer
  Downloading pyppeteer-1.0.2-py3-none-any.whl (83 kB)
Requirement already satisfied: importlib-metadata>=1.4 in c:\users\vladl\anaconda3\lib\sit
e-packages (from Pyppeteer) (4.8.1)
Collecting pyee<9.0.0,>=8.1.0
  Downloading pyee-8.2.2-py2.py3-none-any.whl (12 kB)
Collecting websockets<11.0,>=10.0
  Downloading websockets-10.4-cp39-cp39-win_amd64.whl (101 kB)
Requirement already satisfied: tqdm<5.0.0,>=4.42.1 in c:\users\vladl\anaconda3\lib\site-pa
ckages (from Pyppeteer) (4.62.3)
Requirement already satisfied: certifi>=2021 in c:\users\vladl\anaconda3\lib\site-packages
(from Pyppeteer) (2021.10.8)
Requirement already satisfied: urllib3<2.0.0,>=1.25.8 in c:\users\vladl\anaconda3\lib\site
-packages (from Pyppeteer) (1.26.7)
Requirement already satisfied: appdirs<2.0.0,>=1.4.3 in c:\users\vladl\anaconda3\lib\site-
packages (from Pyppeteer) (1.4.4)
Requirement already satisfied: zipp>=0.5 in c:\users\vladl\anaconda3\lib\site-packages (fr
om importlib-metadata>=1.4->Pyppeteer) (3.6.0)
Requirement already satisfied: colorama in c:\users\vladl\anaconda3\lib\site-packages (fro
m tqdm<5.0.0,>=4.42.1->Pyppeteer) (0.4.4)
Installing collected packages: websockets, pyee, Pyppeteer
Successfully installed Pyppeteer-1.0.2 pyee-8.2.2 websockets-10.4

WARNING: You are using pip version 21.3.1; however, version 23.0.1 is available.
You should consider upgrading via the 'C:\Users\vladl\anaconda3\python.exe -m pip install
--upgrade pip' command.
```

In [27]:

```
!jupyter nbconvert --to webpdf --allow-chromium-download PK1.ipynb
```

```
[NbConvertApp] Converting notebook PK1.ipynb to webpdf
[NbConvertApp] Building PDF
[INFO] Starting Chromium download.

  0%|          | 0.00/137M [00:00<?, ?b/s]
  0%|          | 51.2k/137M [00:00<05:30, 414kb/s]
  0%|          | 102k/137M [00:00<05:36, 407kb/s]
  0%|          | 225k/137M [00:00<03:06, 732kb/s]
  0%|          | 369k/137M [00:00<02:18, 984kb/s]
  1%|          | 696k/137M [00:00<01:17, 1.76Mb/s]
  1%|          | 1.22M/137M [00:00<00:47, 2.89Mb/s]
  1%|1         | 1.96M/137M [00:00<00:31, 4.24Mb/s]
  2%|2         | 2.83M/137M [00:00<00:23, 5.59Mb/s]
  3%|2         | 3.74M/137M [00:00<00:20, 6.64Mb/s]
  3%|3         | 4.69M/137M [00:01<00:17, 7.45Mb/s]
  4%|4         | 5.70M/137M [00:01<00:15, 8.21Mb/s]
  5%|4         | 6.58M/137M [00:01<00:15, 8.37Mb/s]
  6%|5         | 7.54M/137M [00:01<00:14, 8.71Mb/s]
  6%|6         | 8.49M/137M [00:01<00:14, 8.83Mb/s]
  7%|6         | 9.44M/137M [00:01<00:14, 9.01Mb/s]
  8%|7         | 10.5M/137M [00:01<00:13, 9.35Mb/s]
  8%|8         | 11.4M/137M [00:01<00:14, 8.46Mb/s]
  9%|9         | 12.3M/137M [00:01<00:14, 8.62Mb/s]
 10%|9         | 13.5M/137M [00:02<00:13, 9.24Mb/s]
 11%|#         | 14.6M/137M [00:02<00:12, 9.80Mb/s]
 11%|#1        | 15.7M/137M [00:02<00:12, 9.99Mb/s]
 12%|#2        | 16.8M/137M [00:02<00:11, 10.2Mb/s]
 13%|#3        | 17.8M/137M [00:02<00:11, 10.2Mb/s]
 14%|#3        | 18.9M/137M [00:02<00:11, 10.2Mb/s]
 15%|#4        | 20.0M/137M [00:02<00:11, 10.6Mb/s]
 15%|#5        | 21.1M/137M [00:02<00:11, 10.5Mb/s]
 16%|#6        | 22.2M/137M [00:02<00:10, 10.6Mb/s]
 17%|#7        | 23.4M/137M [00:02<00:10, 10.9Mb/s]
 18%|#7        | 24.5M/137M [00:03<00:10, 10.7Mb/s]
 19%|#8        | 25.6M/137M [00:03<00:11, 9.88Mb/s]
 19%|#9        | 26.6M/137M [00:03<00:12, 9.15Mb/s]
 20%|##        | 27.5M/137M [00:03<00:12, 8.86Mb/s]
 21%|##        | 28.4M/137M [00:03<00:12, 9.01Mb/s]
 22%|##1       | 29.6M/137M [00:03<00:10, 9.77Mb/s]
 22%|##2       | 30.8M/137M [00:03<00:10, 10.2Mb/s]
 23%|##3       | 31.8M/137M [00:03<00:10, 10.2Mb/s]
 24%|##3       | 32.8M/137M [00:03<00:10, 9.95Mb/s]
 25%|##4       | 33.9M/137M [00:04<00:10, 9.66Mb/s]
 25%|##5       | 34.8M/137M [00:04<00:11, 9.03Mb/s]
 26%|##6       | 35.7M/137M [00:04<00:12, 8.38Mb/s]
 27%|##6       | 36.6M/137M [00:04<00:12, 8.12Mb/s]
 27%|##7       | 37.4M/137M [00:04<00:13, 7.63Mb/s]
 28%|##7       | 38.2M/137M [00:04<00:13, 7.40Mb/s]
 28%|##8       | 39.0M/137M [00:04<00:13, 7.29Mb/s]
 29%|##8       | 39.7M/137M [00:04<00:13, 7.12Mb/s]
 30%|##9       | 40.4M/137M [00:05<00:16, 5.92Mb/s]
 30%|##9       | 41.1M/137M [00:05<00:15, 6.04Mb/s]
 31%|###       | 41.9M/137M [00:05<00:14, 6.49Mb/s]
 31%|###1      | 42.6M/137M [00:05<00:13, 6.81Mb/s]
 32%|###1      | 43.5M/137M [00:05<00:13, 7.15Mb/s]
 32%|###2      | 44.3M/137M [00:05<00:12, 7.32Mb/s]
 33%|###3      | 45.2M/137M [00:05<00:11, 7.94Mb/s]
 34%|###3      | 46.1M/137M [00:05<00:11, 8.15Mb/s]
 34%|###4      | 47.0M/137M [00:05<00:11, 7.71Mb/s]
 35%|###4      | 47.8M/137M [00:05<00:11, 7.80Mb/s]
 36%|###5      | 48.8M/137M [00:06<00:10, 8.52Mb/s]
 36%|###6      | 49.8M/137M [00:06<00:10, 8.63Mb/s]
 37%|###7      | 50.8M/137M [00:06<00:09, 9.01Mb/s]
 38%|###7      | 51.7M/137M [00:06<00:09, 8.93Mb/s]
 39%|###8      | 52.8M/137M [00:06<00:08, 9.47Mb/s]
 39%|###9      | 54.0M/137M [00:06<00:08, 10.0Mb/s]
 40%|####      | 55.2M/137M [00:06<00:07, 10.8Mb/s]
 41%|####1     | 56.3M/137M [00:06<00:07, 10.9Mb/s]
 42%|####1     | 57.4M/137M [00:06<00:07, 10.1Mb/s]
 43%|####3     | 59.0M/137M [00:07<00:06, 11.7Mb/s]
 44%|####3     | 60.2M/137M [00:07<00:06, 11.5Mb/s]
 45%|####4     | 61.4M/137M [00:07<00:07, 9.63Mb/s]
 46%|####5     | 62.4M/137M [00:07<00:08, 8.59Mb/s]
 46%|####6     | 63.3M/137M [00:07<00:09, 8.07Mb/s]
 47%|####6     | 64.2M/137M [00:07<00:09, 8.07Mb/s]
 48%|####7     | 65.0M/137M [00:07<00:08, 8.19Mb/s]
 48%|####8     | 66.1M/137M [00:07<00:08, 8.74Mb/s]
 49%|####9     | 67.4M/137M [00:08<00:07, 9.66Mb/s]
 50%|#####     | 68.7M/137M [00:08<00:06, 10.5Mb/s]
 51%|#####     | 69.7M/137M [00:08<00:06, 9.65Mb/s]
 52%|#####1    | 70.7M/137M [00:08<00:07, 8.86Mb/s]
 52%|#####2    | 71.6M/137M [00:08<00:08, 7.98Mb/s]
 53%|#####2    | 72.5M/137M [00:08<00:07, 8.07Mb/s]
 54%|#####3    | 73.3M/137M [00:08<00:08, 7.66Mb/s]
 54%|#####4    | 74.1M/137M [00:08<00:08, 7.42Mb/s]
 55%|#####4    | 74.9M/137M [00:08<00:08, 7.22Mb/s]
 55%|#####5    | 75.6M/137M [00:09<00:08, 7.17Mb/s]
 56%|#####5    | 76.3M/137M [00:09<00:08, 7.14Mb/s]
 56%|#####6    | 77.1M/137M [00:09<00:08, 6.94Mb/s]
 57%|#####6    | 77.9M/137M [00:09<00:08, 7.14Mb/s]
 57%|#####7    | 78.6M/137M [00:09<00:08, 6.70Mb/s]
 58%|#####8    | 79.6M/137M [00:09<00:07, 7.59Mb/s]
 59%|#####8    | 80.4M/137M [00:09<00:07, 7.55Mb/s]
 60%|#####9    | 81.5M/137M [00:09<00:06, 8.42Mb/s]
 60%|######    | 82.3M/137M [00:09<00:06, 8.24Mb/s]
 61%|######    | 83.2M/137M [00:10<00:06, 8.34Mb/s]
 61%|######1   | 84.1M/137M [00:10<00:07, 7.44Mb/s]
 62%|######1   | 84.8M/137M [00:10<00:07, 6.74Mb/s]
 63%|######2   | 85.8M/137M [00:10<00:06, 7.31Mb/s]
 63%|######3   | 86.9M/137M [00:10<00:06, 8.06Mb/s]
 64%|######4   | 87.8M/137M [00:10<00:05, 8.48Mb/s]
 65%|######4   | 88.8M/137M [00:10<00:05, 8.68Mb/s]
 66%|######5   | 89.8M/137M [00:10<00:05, 9.05Mb/s]
 66%|######6   | 90.9M/137M [00:10<00:04, 9.47Mb/s]
 67%|######7   | 92.0M/137M [00:11<00:04, 10.0Mb/s]
 68%|######7   | 93.0M/137M [00:11<00:04, 9.86Mb/s]
 69%|######8   | 94.0M/137M [00:11<00:04, 9.49Mb/s]
 69%|######9   | 95.0M/137M [00:11<00:04, 9.02Mb/s]
 70%|#######   | 96.1M/137M [00:11<00:04, 9.57Mb/s]
 71%|#######   | 97.1M/137M [00:11<00:04, 9.63Mb/s]
 72%|#######1  | 98.1M/137M [00:11<00:03, 9.82Mb/s]
 72%|#######2  | 99.1M/137M [00:11<00:03, 9.46Mb/s]
 73%|#######3  | 100M/137M [00:11<00:04, 8.16Mb/s]
 74%|#######4  | 102M/137M [00:12<00:03, 9.68Mb/s]
 75%|#######4  | 103M/137M [00:12<00:03, 10.0Mb/s]
 76%|#######5  | 104M/137M [00:12<00:03, 9.85Mb/s]
 76%|#######6  | 105M/137M [00:12<00:03, 9.82Mb/s]
 77%|#######7  | 106M/137M [00:12<00:03, 9.73Mb/s]
 78%|#######7  | 107M/137M [00:12<00:03, 9.45Mb/s]
 79%|#######8  | 108M/137M [00:12<00:03, 9.25Mb/s]
 79%|#######9  | 109M/137M [00:12<00:02, 9.51Mb/s]
 80%|########  | 110M/137M [00:12<00:02, 9.61Mb/s]
 81%|########  | 111M/137M [00:13<00:02, 9.39Mb/s]
 82%|########1 | 112M/137M [00:13<00:02, 9.69Mb/s]
 82%|########2 | 113M/137M [00:13<00:02, 9.05Mb/s]
 83%|########3 | 114M/137M [00:13<00:02, 10.3Mb/s]
 84%|########4 | 115M/137M [00:13<00:02, 9.69Mb/s]
 85%|########4 | 116M/137M [00:13<00:02, 9.30Mb/s]
 86%|########5 | 117M/137M [00:13<00:02, 9.35Mb/s]
 86%|########6 | 118M/137M [00:13<00:02, 9.29Mb/s]
 87%|########6 | 119M/137M [00:13<00:01, 9.28Mb/s]
 88%|########7 | 120M/137M [00:14<00:01, 9.29Mb/s]
 88%|########8 | 121M/137M [00:14<00:01, 9.42Mb/s]
 89%|########8 | 122M/137M [00:14<00:01, 9.35Mb/s]
 90%|########9 | 123M/137M [00:14<00:01, 9.28Mb/s]
 90%|######### | 124M/137M [00:14<00:01, 9.67Mb/s]
 91%|#########1| 125M/137M [00:14<00:01, 9.52Mb/s]
 92%|#########1| 126M/137M [00:14<00:01, 9.44Mb/s]
 93%|#########2| 127M/137M [00:14<00:01, 8.06Mb/s]
 94%|#########3| 129M/137M [00:14<00:00, 10.7Mb/s]
 95%|#########4| 130M/137M [00:15<00:00, 10.9Mb/s]
 96%|#########5| 131M/137M [00:15<00:00, 10.9Mb/s]
 97%|#########6| 132M/137M [00:15<00:00, 11.4Mb/s]
 97%|#########7| 133M/137M [00:15<00:00, 10.4Mb/s]
 98%|#########8| 134M/137M [00:15<00:00, 8.20Mb/s]
 99%|#########8| 135M/137M [00:15<00:00, 8.31Mb/s]
100%|#########9| 136M/137M [00:15<00:00, 8.75Mb/s]
```

```
100%|##########| 137M/137M [00:15<00:00, 8.65Mb/s]
[INFO] Beginning extraction
[INFO] Chromium extracted to: C:\Users\vladl\AppData\Local\pyppeteer\pyppeteer\local-chrom
ium\588429
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 208249 bytes to PK1.pdf
```

In [ ]:

```

```