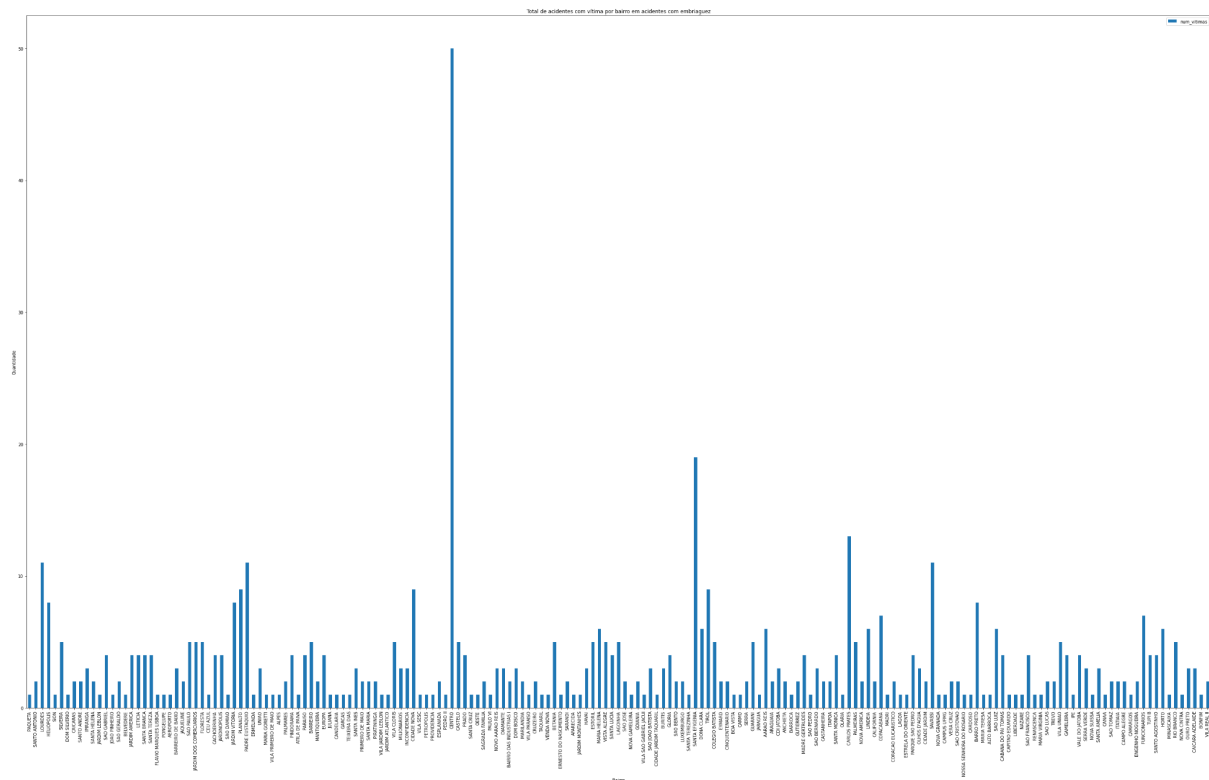


ESD21 - Data Engineering (Big Data & Analytics)

Avaliação final.

Total de acidentes com vítima por bairro em acidentes com embriaguez.



Comando python para ser executado com spark

```
# -*- coding: utf-8 -*-
from pyspark.sql import SparkSession
from pyspark.sql import Row
import matplotlib.pyplot as plt

spark = SparkSession \
    .builder \
    .appName("Total de acidentes com vitima por bairro em acidentes com embriaguez") \
    .enableHiveSupport() \
    .getOrCreate()

spark.sql("USE trabalhoFinal")
```

```

df = spark.sql("SELECT log.nome_bairro, count(*) AS num_vitimas FROM "
+
+           "si_env AS env " +
+           "INNER JOIN si_log_2019 AS log " +
+           "ON env.num_boletim = log.n_boletim WHERE
env.Embreagues = 'SIM'" +
+           "GROUP BY log.nome_bairro")

df.coalesce(1).write \
.option("header", "true") \
.format("csv") \
.save("/user/vagrant/atividades/resultados/atividades01")

```

colab

```

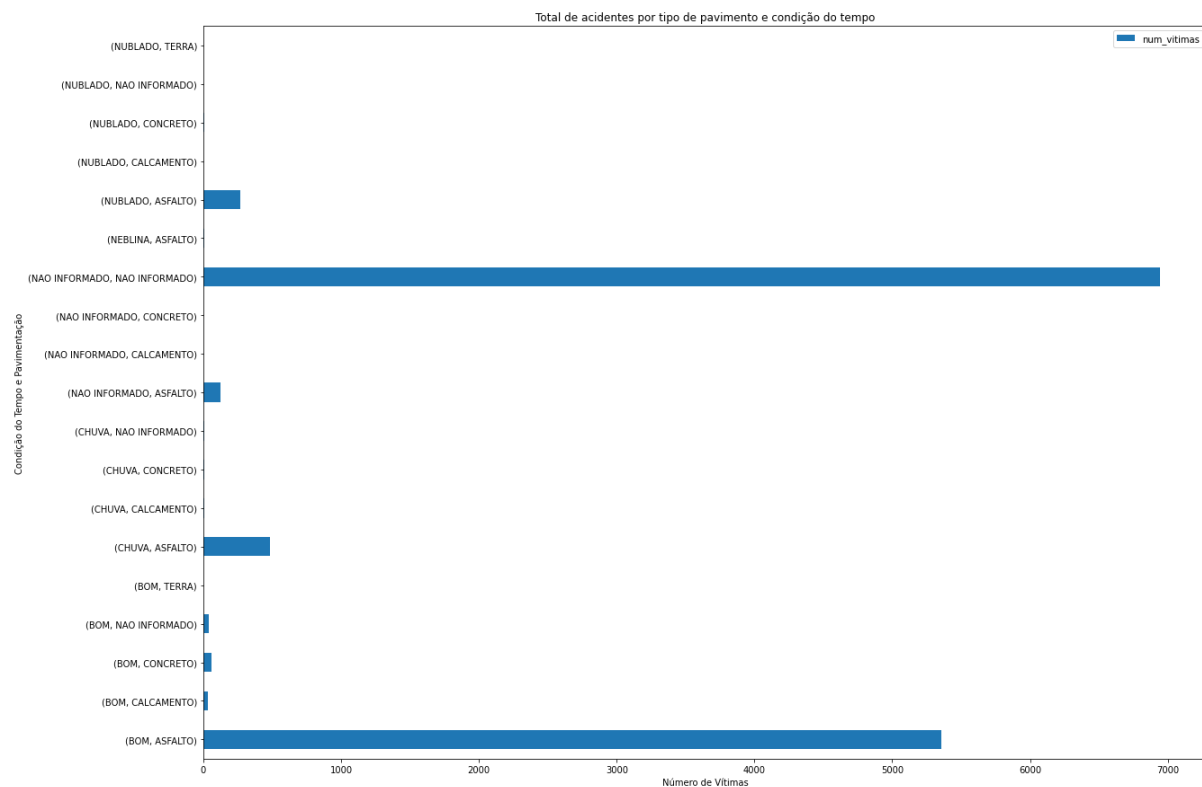
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("/content/dados/atividade01.csv", sep=',', header=0,
encoding='latin_1')
df.plot.bar(x='nome_bairro', y='num_vitimas', figsize=(50, 30))

plt.title('Total de acidentes com vítima por bairro em acidentes com
embriaguez')
plt.xlabel('Bairro')
plt.ylabel('Quantidade')

```

Total de acidentes por tipo de pavimento e condição do tempo.



Comando python para ser executado com spark

```
# -*- coding: utf-8 -*-
from pyspark.sql import SparkSession
from pyspark.sql import Row
import matplotlib.pyplot as plt

spark = SparkSession \
    .builder \
    .appName("Total de acidentes com vitima por bairro em acidentes com embriaguez") \
    .enableHiveSupport() \
    .getOrCreate()

spark.sql("USE trabalhoFinal")

df = spark.sql("SELECT desc_tempo, pavimento, COUNT(*) AS num_vitimas
FROM " +
               "si_bol_2019 GROUP BY desc_tempo, pavimento")

df.coalesce(1).write \
    .option("header", "true") \
    .format("csv") \
    .save("/user/vagrant/atividades/resultados/atividades02")
```

Colab

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

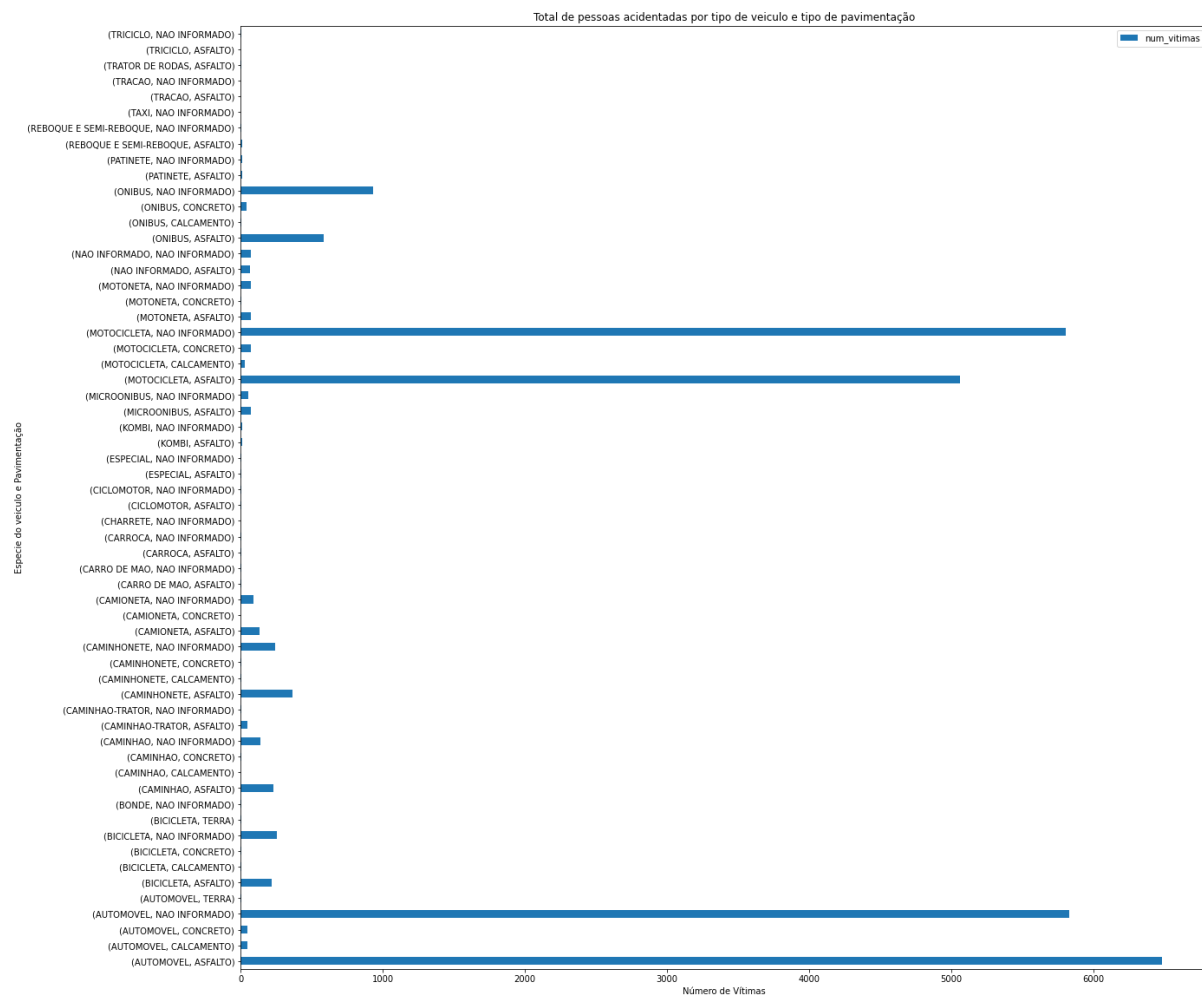
df = pd.read_csv("/content/dados/atividade02.csv", sep=',', header=0,
encoding='latin-1')

df.num_vitimas=pd.to_numeric(df.num_vitimas)

df[['desc_tempo','pavimento','num_vitimas']].groupby(['desc_tempo','pavimento']).sum().plot.barh(figsize=(20,15))

plt.title('Total de acidentes por tipo de pavimento e condição do tempo')
plt.ylabel('Condição do Tempo e Pavimentação')
plt.xlabel('Número de Vítimas')
```

Total de pessoas acidentadas por tipo de veículo e tipo de pavimentação.



Comando python para ser executado com spark

```
# -*- coding: utf-8 -*-

from pyspark.sql import SparkSession
from pyspark.sql import Row
import matplotlib.pyplot as plt

spark = SparkSession \
    .builder \
    .appName("Total de acidentes com vitima por bairro em acidentes com embriaguez") \
    .enableHiveSupport() \
    .getOrCreate()

spark.sql("USE trabalhoFinal")

df = spark.sql("SELECT env.especie_veiculo, bol.pavimento, COUNT(*) AS num_vitimas FROM " +
               "si_bol_2019 AS bol " +
```

```

            "INNER JOIN si_env AS env ON bol.n_boletim =
env.num_boletim " +
            "GROUP BY env.especie_veiculo, bol.pavimento")

df.coalesce(1).write \
.option("header", "true") \
.format("csv") \
.save("/user/vagrant/atividades/resultados/atividades03")

```

colab

```

import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("/content/dados/atividade03.csv", sep=',', header=0,
encoding='latin_1')

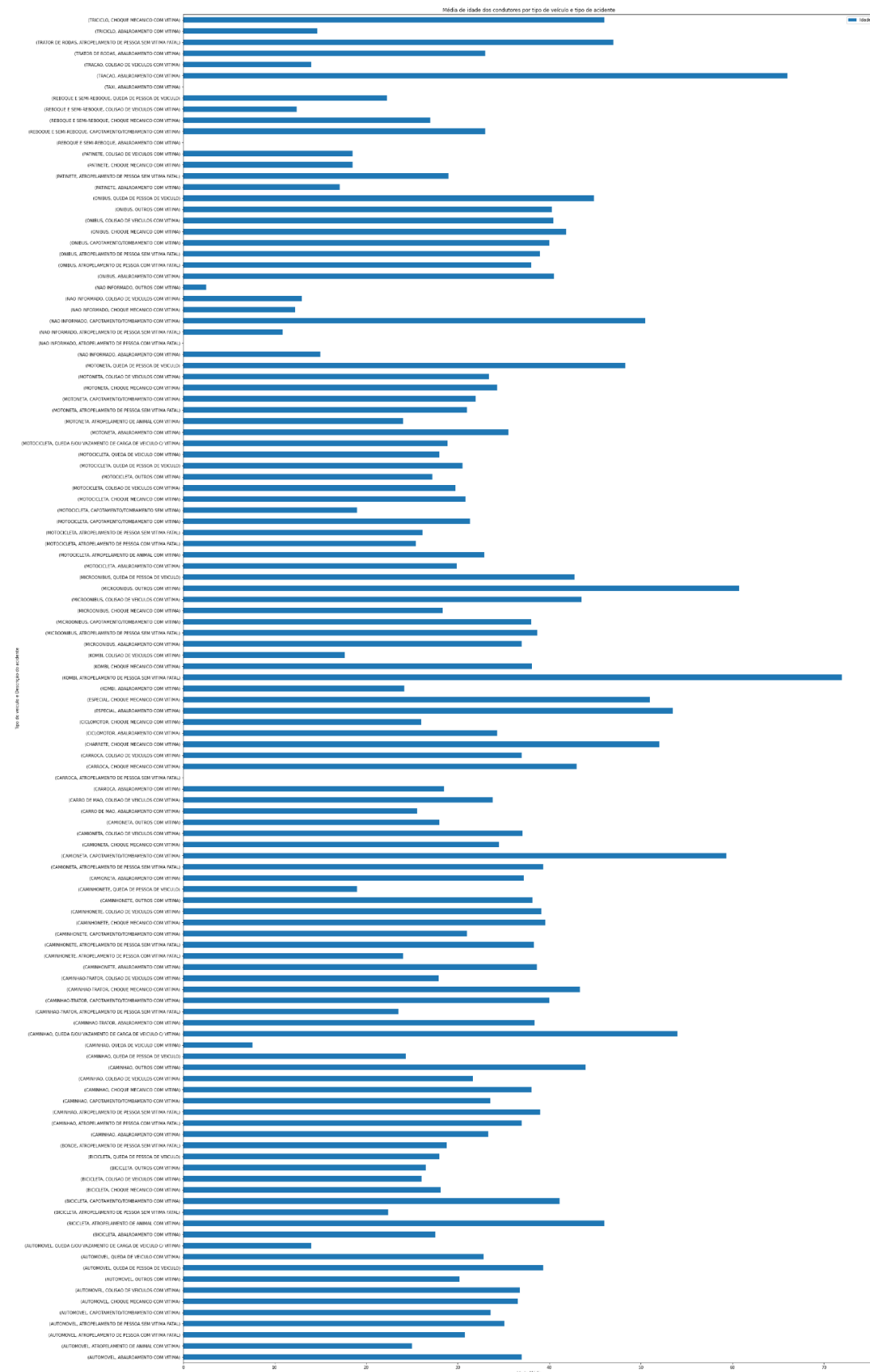
df.num_vitimas=pd.to_numeric(df.num_vitimas)

df[['especie_veiculo','pavimento','num_vitimas']].groupby(['especie_vei
culo','pavimento']).sum().plot.barh(figsize=(20,20))

plt.title('Total de pessoas acidentadas por tipo de veiculo e tipo de
pavimentação')
plt.ylabel('Especie do veiculo e Pavimentação')
plt.xlabel('Número de Vítimas')

```

Média de idade dos condutores por tipo de veículo e tipo de acidente.



Comando python para ser executado com spark

```
# -*- coding: utf-8 -*-
```

```

from pyspark.sql import SparkSession
from pyspark.sql import Row
import matplotlib.pyplot as plt

spark = SparkSession \
    .builder \
    .appName("Total de acidentes com vitima por bairro em acidentes com embriaguez") \
    .enableHiveSupport() \
    .getOrCreate()

spark.sql("USE trabalhoFinal")

df = spark.sql("SELECT env.especie_veiculo, bol.desc_tipo_acidente, AVG(env.Idade) AS Idade from " +
               "si_bol_2019 AS bol " +
               "INNER JOIN si_env AS env ON bol.n_boletim = env.num_boletim " +
               "GROUP BY env.especie_veiculo, bol.desc_tipo_acidente")
df.coalesce(1).write \
    .option("header", "true") \
    .format("csv") \
    .save("/user/vagrant/atividades/resultados/atividades_04")

```

colab

```

import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("/content/dados/atividade04.csv", sep=',', header=0, encoding='latin_1')

df.Idade=pd.to_numeric(df.Idade)

df[['especie_veiculo','desc_tipo_acidente','Idade']].groupby(['especie_veiculo','desc_tipo_acidente']).sum().plot.barh(figsize=(30,60))

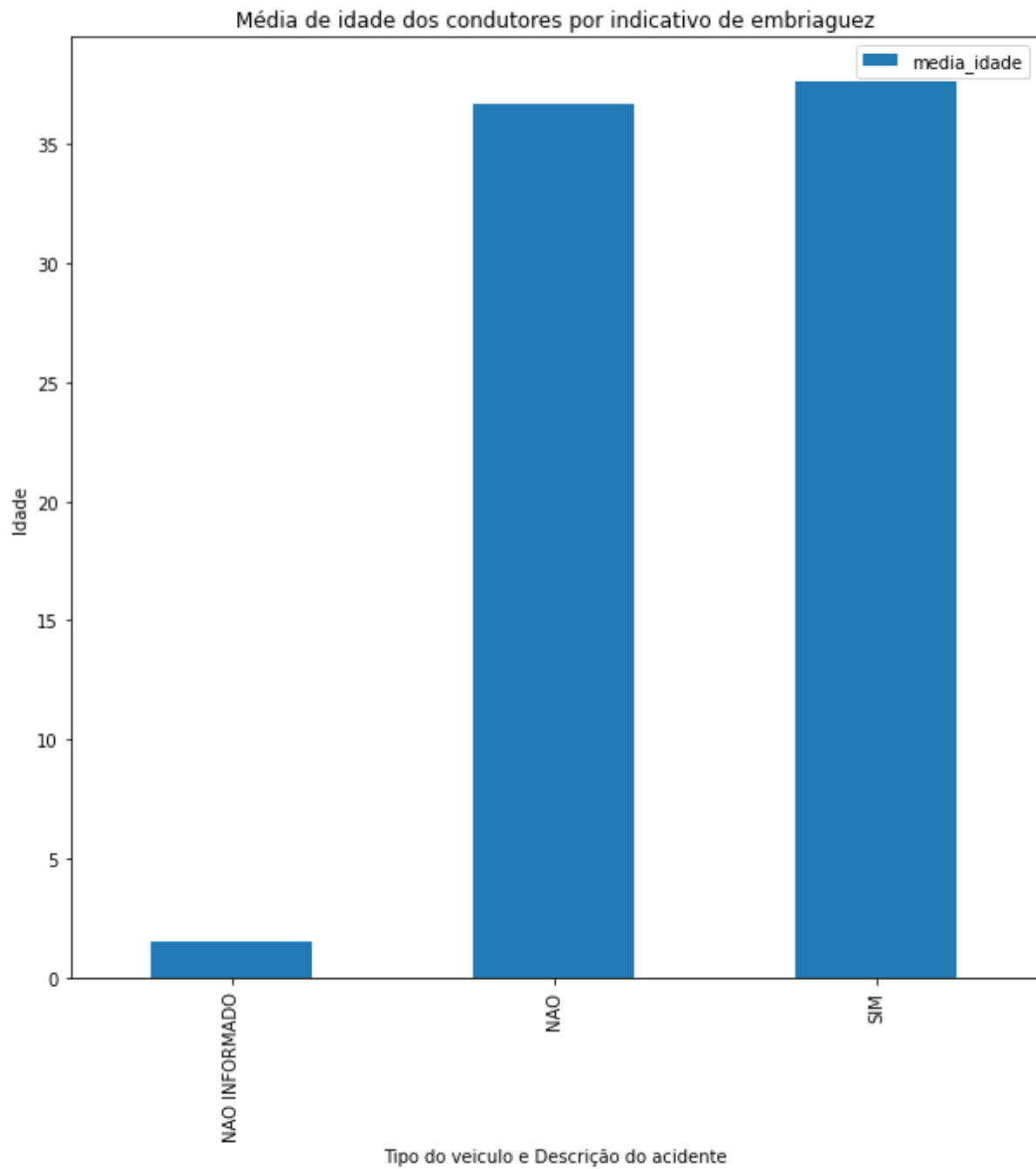
plt.title('Média de idade dos condutores por tipo de veículo e tipo de acidente')
plt.ylabel('Tipo do veiculo e Descrição do acidente')

```



```
plt.xlabel('Idade Média')
```

Média de idade dos condutores por indicativo de embriaguez.



Comando python para ser executado com spark

```
# -*- coding: utf-8 -*-  
from pyspark.sql import SparkSession  
from pyspark.sql import Row  
import matplotlib.pyplot as plt  
  
spark = SparkSession \
```

```

        .builder \
        .appName("Total de acidentes com vitima por bairro em acidentes com
embriaguez") \
        .enableHiveSupport() \
        .getOrCreate()

spark.sql("USE trabalhoFinal")

df = spark.sql("SELECT Embreagues, AVG(Idade) AS media_idade FROM
si_env GROUP BY Embreagues")

df.coalesce(1).write \
.option("header", "true") \
.format("csv") \
.save("/user/vagrant/atividades/resultados/atividades05")

```

colab

```

import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("/content/dados/atividade05.csv", sep=',', header=0,
encoding='latin_1')

df.plot.bar(x='Embreagues', figsize=(10,10))

plt.title('Média de idade dos condutores por indicativo de embriaguez')
plt.xlabel('Tipo do veiculo e Descrição do acidente')
plt.ylabel('Idade')

```