AED-CPA-tabelas

October 27, 2024

1 CPA Tabelas

1.1 Carregamento de Dados e Impotação de Bibliotecas

[2]:
```
import pandas as pd
import numpy as np
import os
import plotly.express as px
```

[1]:
```
xls_path = 'C:/Users/Inteli/Documents/GitHub/2024-2B-T10-SI08-G01/src/dados/
 ↪CPA_Tabelas.xlsx'

folder_path = 'C:/Users/Inteli/Documents/GitHub/2024-2B-T10-SI08-G01/src/dados'

# arrega o arquivo XLSX
xls = pd.ExcelFile(xls_path)

# itera cada aba e salva como csv
for sheet_name in xls.sheet_names:
    df = pd.read_excel(xls, sheet_name=sheet_name)
    # salva cada aba como csv usando o nome da aba
    df.to_csv(os.path.join(folder_path, f'{sheet_name}.csv'), index=False)
```

[2]:
```
dataframes = []

for filename in os.listdir(folder_path):
    if filename.endswith('.csv'):  # verifica se o arquivo é um CSV
        file_path = os.path.join(folder_path, filename)
        df = pd.read_csv(file_path)
        dataframes.append(df)

# concatena todos os dfs em um único
combined_df = pd.concat(dataframes, ignore_index=True)

# salva o df em um novo csv
combined_df.to_csv('C:/Users/Inteli/Documents/GitHub/2024-2B-T10-SI08-G01/src/
 ↪dados/CPA_MOVIMENTO_COMPLETO.csv', index=False)
```

1.2 Análise Exploratória Inicial

[3]:
```
df_movimento = pd.read_csv('C:/Users/Inteli/Documents/GitHub/
↪2024-2B-T10-SI08-G01/src/dados/CPA_MOVIMENTO_COMPLETO.csv')

df_movimento.head()
```

```
C:\Users\Inteli\AppData\Local\Temp\ipykernel_8992\3158990088.py:1: DtypeWarning:
Columns (33,34,40,43,47,50,51) have mixed types. Specify dtype option on import
or set low_memory=False.
  df_movimento = pd.read_csv('C:/Users/Inteli/Documents/GitHub/2024-2B-T10-SI08-
G01/src/dados/CPA_MOVIMENTO_COMPLETO.csv')
```

[3]:

	ID_MOVIMENTO	CD_MOVIMENTO	TX_ESTACAO_ORIGEM	ID_ESTACAO_ORIGEM	\
0	177.0	80.0	BFU	9.0	
1	179.0	81.0	BFU	9.0	
2	197.0	90.0	BFU	9.0	
3	199.0	91.0	BFU	9.0	
4	211.0	96.0	BFU	9.0	

	TX_ESTACAO_DESTINO	ID_ESTACAO_DESTINO	CD_HORARIO_FK	\
0	DMO	29.0	39.0	
1	DMO	29.0	40.0	
2	GMC	38.0	39.0	
3	GMC	38.0	40.0	
4	IPV	48.0	39.0	

	CD_HORARIO_PARTIDA_FK	CD_HORARIO_CHEGADA_FK	CD_HORARIO_AJUSTE_FK	…	\
0	39.0	39.0	NaN	…	
1	40.0	40.0	NaN	…	
2	39.0	41.0	NaN	…	
3	40.0	42.0	NaN	…	
4	39.0	43.0	NaN	…	

	CD_TRECHO	NR_SEQ_ESTACAOLINHA	ID_REGISTRO	ID_CODUSUARIO	DT_ATUALIZA	\
0	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	
2	NaN	NaN	NaN	NaN	NaN	
3	NaN	NaN	NaN	NaN	NaN	
4	NaN	NaN	NaN	NaN	NaN	

	FL_FIM_TRECHO	ID_LINHA_FK	TX_DESCRICAO	DT_INICIO	DT_FIM
0	NaN	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN

[5 rows x 53 columns]

[7]:
```
df_movimento.describe()
```

[7]:

	ID_MOVIMENTO	CD_MOVIMENTO	ID_ESTACAO_ORIGEM	ID_ESTACAO_DESTINO	\
count	1.219224e+06	1.219224e+06	1.219224e+06	1.219224e+06	
mean	1.069764e+06	1.035475e+05	4.604292e+01	4.644101e+01	
std	7.645512e+05	6.121995e+04	2.938355e+01	2.943122e+01	
min	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	
25%	3.048068e+05	5.144400e+04	1.300000e+01	1.300000e+01	
50%	1.135146e+06	1.013580e+05	4.400000e+01	4.600000e+01	
75%	1.439951e+06	1.515430e+05	6.800000e+01	6.900000e+01	
max	2.303563e+06	2.263910e+05	1.060000e+02	1.060000e+02	

	CD_HORARIO_FK	CD_HORARIO_PARTIDA_FK	CD_HORARIO_CHEGADA_FK	\
count	1.219224e+06	1.219224e+06	1.219224e+06	
mean	4.143906e+01	4.245535e+01	4.372008e+01	
std	2.317713e+01	2.325759e+01	2.327634e+01	
min	1.000000e+00	1.000000e+00	1.000000e+00	
25%	2.100000e+01	2.200000e+01	2.300000e+01	
50%	4.200000e+01	4.300000e+01	4.400000e+01	
75%	6.100000e+01	6.200000e+01	6.300000e+01	
max	8.800000e+01	9.300000e+01	9.400000e+01	

	CD_HORARIO_AJUSTE_FK	NR_AJUSTE_MINUTO	NR_ACUMULADO_MINUTO	…	\
count	0.0	1.219224e+06	0.0	…	
mean	NaN	2.053627e+01	NaN	…	
std	NaN	1.506854e+01	NaN	…	
min	NaN	1.000000e+00	NaN	…	
25%	NaN	7.000000e+00	NaN	…	
50%	NaN	1.600000e+01	NaN	…	
75%	NaN	3.000000e+01	NaN	…	
max	NaN	6.500000e+01	NaN	…	

	TEMPO_TRANSF	FL_ESTACAO	ID_ESTACAO	ID_LINHA	NR_SEQ_ESTACAOLINHA	\
count	28.000000	105.000000	99.000000	99.000000	99.000000	
mean	14.428571	0.942857	52.848485	3.585859	89.646465	
std	1.708987	0.233229	30.062983	1.767104	50.311542	
min	12.000000	0.000000	1.000000	1.000000	10.000000	
25%	12.000000	1.000000	27.500000	2.000000	50.000000	
50%	15.000000	1.000000	53.000000	4.000000	80.000000	
75%	15.000000	1.000000	78.500000	5.000000	130.000000	
max	17.000000	1.000000	106.000000	6.000000	200.000000	

	ID_REGISTRO	ID_CODUSUARIO	FL_FIM_TRECHO	ID_LINHA_FK	DT_FIM
count	99.000000	107.0	99.000000	8.000000	0.0
mean	50.000000	2113.0	0.161616	3.125000	NaN

```
std      28.722813           0.0       0.369972     1.807722     NaN
min       1.000000        2113.0       0.000000     1.000000     NaN
25%      25.500000        2113.0       0.000000     1.750000     NaN
50%      50.000000        2113.0       0.000000     3.000000     NaN
75%      74.500000        2113.0       0.000000     4.250000     NaN
max      99.000000        2113.0       1.000000     6.000000     NaN

[8 rows x 41 columns]
```

[8]:
```
df_movimento.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1219436 entries, 0 to 1219435
Data columns (total 53 columns):
 #   Column                  Non-Null Count    Dtype
---  ------                  --------------    -----
 0   ID_MOVIMENTO            1219224 non-null  float64
 1   CD_MOVIMENTO            1219224 non-null  float64
 2   TX_ESTACAO_ORIGEM       1219224 non-null  object
 3   ID_ESTACAO_ORIGEM       1219224 non-null  float64
 4   TX_ESTACAO_DESTINO      1219224 non-null  object
 5   ID_ESTACAO_DESTINO      1219224 non-null  float64
 6   CD_HORARIO_FK           1219224 non-null  float64
 7   CD_HORARIO_PARTIDA_FK   1219224 non-null  float64
 8   CD_HORARIO_CHEGADA_FK   1219224 non-null  float64
 9   CD_HORARIO_AJUSTE_FK    0 non-null        float64
 10  NR_AJUSTE_MINUTO        1219224 non-null  float64
 11  NR_ACUMULADO_MINUTO     0 non-null        float64
 12  NR_PASSAGEIRO           1219224 non-null  float64
 13  ID_TRECHO_ORIGEM        1219224 non-null  object
 14  ID_TRECHO_DESTINO       1219224 non-null  object
 15  FL_SENTIDO              1219224 non-null  float64
 16  ID_LINHA_ORIGEM         1219224 non-null  float64
 17  ID_LINHA_DESTINO        1219224 non-null  float64
 18  CD_TIPO                 1219224 non-null  float64
 19  CD_HOR_PARTIDA_FK       0 non-null        float64
 20  CD_HOR_CHEGADA_FK       0 non-null        float64
 21  NR_TEMPO_ESPERA         0 non-null        float64
 22  NR_AJUSTE_KM            0 non-null        float64
 23  NR_ACUMULADO_KM         0 non-null        float64
 24  NR_PASSAGEIRO_ORIGINAL  0 non-null        float64
 25  CD_MOV                  0 non-null        float64
 26  CD_OD                   0 non-null        float64
 27  FL_ORIGINAL             0 non-null        float64
 28  FL_ACERTO               0 non-null        float64
 29  TX_DIAS                 1219224 non-null  object
 30  ID_ANTERIOR             1219224 non-null  float64
 31  LIN_CD_LINHA            105 non-null      float64
```

```
 32  CD_ESTACAO              104 non-null      float64
 33  TX_PREFIXO              105 non-null      object
 34  TX_ESTACAO              204 non-null      object
 35  TRANSF_ID_LINHA         28 non-null       float64
 36  TRANSF_ID_EST_DEST      28 non-null       float64
 37  TEMPO_PERC              105 non-null      float64
 38  TEMPO_TRANSF            28 non-null       float64
 39  FL_ESTACAO              105 non-null      float64
 40  TX_MODAL                105 non-null      object
 41  ID_ESTACAO              99 non-null       float64
 42  ID_LINHA                99 non-null       float64
 43  CD_TRECHO               107 non-null      object
 44  NR_SEQ_ESTACAOLINHA     99 non-null       float64
 45  ID_REGISTRO             99 non-null       float64
 46  ID_CODUSUARIO           107 non-null      float64
 47  DT_ATUALIZA             107 non-null      object
 48  FL_FIM_TRECHO           99 non-null       float64
 49  ID_LINHA_FK             8 non-null        float64
 50  TX_DESCRICAO            8 non-null        object
 51  DT_INICIO               8 non-null        object
 52  DT_FIM                  0 non-null        float64
dtypes: float64(41), object(12)
memory usage: 493.1+ MB
```

[12]:
```
df_movimento.isnull().sum()
```

[12]:
```
ID_MOVIMENTO                212
CD_MOVIMENTO                212
TX_ESTACAO_ORIGEM           212
ID_ESTACAO_ORIGEM           212
TX_ESTACAO_DESTINO          212
ID_ESTACAO_DESTINO          212
CD_HORARIO_FK               212
CD_HORARIO_PARTIDA_FK       212
CD_HORARIO_CHEGADA_FK       212
CD_HORARIO_AJUSTE_FK    1219436
NR_AJUSTE_MINUTO            212
NR_ACUMULADO_MINUTO     1219436
NR_PASSAGEIRO               212
ID_TRECHO_ORIGEM            212
ID_TRECHO_DESTINO           212
FL_SENTIDO                  212
ID_LINHA_ORIGEM             212
ID_LINHA_DESTINO            212
CD_TIPO                     212
CD_HOR_PARTIDA_FK       1219436
CD_HOR_CHEGADA_FK       1219436
```

```
NR_TEMPO_ESPERA           1219436
NR_AJUSTE_KM              1219436
NR_ACUMULADO_KM           1219436
NR_PASSAGEIRO_ORIGINAL    1219436
CD_MOV                    1219436
CD_OD                     1219436
FL_ORIGINAL               1219436
FL_ACERTO                 1219436
TX_DIAS                       212
ID_ANTERIOR                   212
LIN_CD_LINHA              1219331
CD_ESTACAO                1219332
TX_PREFIXO                1219331
TX_ESTACAO                1219232
TRANSF_ID_LINHA           1219408
TRANSF_ID_EST_DEST        1219408
TEMPO_PERC                1219331
TEMPO_TRANSF              1219408
FL_ESTACAO                1219331
TX_MODAL                  1219331
ID_ESTACAO                1219337
ID_LINHA                  1219337
CD_TRECHO                 1219329
NR_SEQ_ESTACAOLINHA       1219337
ID_REGISTRO               1219337
ID_CODUSUARIO             1219329
DT_ATUALIZA               1219329
FL_FIM_TRECHO             1219337
ID_LINHA_FK               1219428
TX_DESCRICAO              1219428
DT_INICIO                 1219428
DT_FIM                    1219436
dtype: int64
```

1.3 Limpeza de Dados

[4]:
```
df_m_limpo_nulos = df_movimento.dropna(axis=1, how='all')

df_m_limpo_nulos
```

[4]:

	ID_MOVIMENTO	CD_MOVIMENTO	TX_ESTACAO_ORIGEM	ID_ESTACAO_ORIGEM	\
0	177.0	80.0	BFU	9.0	
1	179.0	81.0	BFU	9.0	
2	197.0	90.0	BFU	9.0	
3	199.0	91.0	BFU	9.0	
4	211.0	96.0	BFU	9.0	
…	…	…	…	…	

```
1219431           NaN           NaN               NaN               NaN
1219432           NaN           NaN               NaN               NaN
1219433           NaN           NaN               NaN               NaN
1219434           NaN           NaN               NaN               NaN
1219435           NaN           NaN               NaN               NaN

        TX_ESTACAO_DESTINO  ID_ESTACAO_DESTINO  CD_HORARIO_FK  \
0                      DMO                29.0           39.0
1                      DMO                29.0           40.0
2                      GMC                38.0           39.0
3                      GMC                38.0           40.0
4                      IPV                48.0           39.0
…                      …                  …              …
1219431                NaN                NaN            NaN
1219432                NaN                NaN            NaN
1219433                NaN                NaN            NaN
1219434                NaN                NaN            NaN
1219435                NaN                NaN            NaN

         CD_HORARIO_PARTIDA_FK  CD_HORARIO_CHEGADA_FK  NR_AJUSTE_MINUTO  …  \
0                         39.0                   39.0               8.0  …
1                         40.0                   40.0               8.0  …
2                         39.0                   41.0              28.0  …
3                         40.0                   42.0              28.0  …
4                         39.0                   43.0              56.0  …
…                         …                      …                 …  …
1219431                    NaN                    NaN               NaN  …
1219432                    NaN                    NaN               NaN  …
1219433                    NaN                    NaN               NaN  …
1219434                    NaN                    NaN               NaN  …
1219435                    NaN                    NaN               NaN  …

         ID_LINHA CD_TRECHO NR_SEQ_ESTACAOLINHA  ID_REGISTRO  ID_CODUSUARIO  \
0             NaN       NaN                 NaN          NaN            NaN
1             NaN       NaN                 NaN          NaN            NaN
2             NaN       NaN                 NaN          NaN            NaN
3             NaN       NaN                 NaN          NaN            NaN
4             NaN       NaN                 NaN          NaN            NaN
…             …         …                   …            …              …
1219431       NaN        E1                 NaN          NaN         2113.0
1219432       NaN        E2                 NaN          NaN         2113.0
1219433       NaN        F1                 NaN          NaN         2113.0
1219434       NaN        B1                 NaN          NaN         2113.0
1219435       NaN        C1                 NaN          NaN         2113.0

                DT_ATUALIZA  FL_FIM_TRECHO ID_LINHA_FK  TX_DESCRICAO  \
0                       NaN            NaN         NaN           NaN
```

```
1                      NaN            NaN         NaN           NaN
2                      NaN            NaN         NaN           NaN
3                      NaN            NaN         NaN           NaN
4                      NaN            NaN         NaN           NaN
…                      …              …           …             …
1219431  2022-07-06 08:07:51            NaN         3.0     LUZ - GUA
1219432  2022-07-06 08:07:51            NaN         3.0     GUA - EST
1219433  2022-07-06 08:07:51            NaN         4.0     BAS - CVN
1219434  2022-07-06 08:07:51            NaN         5.0     JPR - IPV
1219435  2022-07-06 08:07:51            NaN         6.0     OSA - GRA

          DT_INICIO
0               NaN
1               NaN
2               NaN
3               NaN
4               NaN
…               …
1219431  2003-01-01
1219432  2003-01-01
1219433  2003-01-01
1219434  2003-01-01
1219435  2003-01-01

[1219436 rows x 40 columns]
```

[5]:
```
df_m_limpo = df_m_limpo_nulos[:-212]

df_m_limpo
```

[5]:
```
          ID_MOVIMENTO  CD_MOVIMENTO TX_ESTACAO_ORIGEM  ID_ESTACAO_ORIGEM  \
0                177.0          80.0               BFU                9.0
1                179.0          81.0               BFU                9.0
2                197.0          90.0               BFU                9.0
3                199.0          91.0               BFU                9.0
4                211.0          96.0               BFU                9.0
…                …              …                  …                  …
1219219     2267713.0      141785.0               SOC               91.0
1219220     2267714.0      141786.0               SOC               91.0
1219221     2267715.0      141787.0               SOC               91.0
1219222     2268970.0      142456.0               SOC               91.0
1219223     2268972.0      142457.0               SOC               91.0

         TX_ESTACAO_DESTINO  ID_ESTACAO_DESTINO  CD_HORARIO_FK  \
0                       DMO                29.0           39.0
1                       DMO                29.0           40.0
2                       GMC                38.0           39.0
```

```
3                    GMC               38.0          40.0
4                    IPV               48.0          39.0
…                    …                 …             …
1219219              CJD               23.0          12.0
1219220              USP               24.0          11.0
1219221              USP               24.0          12.0
1219222              PAL               79.0          48.0
1219223              PAL               79.0          47.0

         CD_HORARIO_PARTIDA_FK  CD_HORARIO_CHEGADA_FK  NR_AJUSTE_MINUTO  …  \
0                         39.0                   39.0               8.0  …
1                         40.0                   40.0               8.0  …
2                         39.0                   41.0              28.0  …
3                         40.0                   42.0              28.0  …
4                         39.0                   43.0              56.0  …
…                         …                      …                 …  …
1219219                   12.0                   12.0              14.0  …
1219220                   11.0                   12.0              19.0  …
1219221                   12.0                   13.0              19.0  …
1219222                   48.0                   49.0              27.0  …
1219223                   47.0                   48.0              27.0  …

         ID_LINHA CD_TRECHO NR_SEQ_ESTACAOLINHA  ID_REGISTRO  ID_CODUSUARIO  \
0             NaN       NaN                  NaN          NaN            NaN
1             NaN       NaN                  NaN          NaN            NaN
2             NaN       NaN                  NaN          NaN            NaN
3             NaN       NaN                  NaN          NaN            NaN
4             NaN       NaN                  NaN          NaN            NaN
…             …         …                    …            …              …
1219219       NaN       NaN                  NaN          NaN            NaN
1219220       NaN       NaN                  NaN          NaN            NaN
1219221       NaN       NaN                  NaN          NaN            NaN
1219222       NaN       NaN                  NaN          NaN            NaN
1219223       NaN       NaN                  NaN          NaN            NaN

         DT_ATUALIZA  FL_FIM_TRECHO ID_LINHA_FK  TX_DESCRICAO  DT_INICIO
0                NaN            NaN         NaN           NaN        NaN
1                NaN            NaN         NaN           NaN        NaN
2                NaN            NaN         NaN           NaN        NaN
3                NaN            NaN         NaN           NaN        NaN
4                NaN            NaN         NaN           NaN        NaN
…                …              …           …             …          …
1219219          NaN            NaN         NaN           NaN        NaN
1219220          NaN            NaN         NaN           NaN        NaN
1219221          NaN            NaN         NaN           NaN        NaN
1219222          NaN            NaN         NaN           NaN        NaN
1219223          NaN            NaN         NaN           NaN        NaN
```

```
[1219224 rows x 40 columns]
```

as ultimas 212 linhas foram desconsideradas por possuírem majoritariamente valores nulos

1.4 Análise Exploratória Avançada

1.4.1 Relação entre variáveis numéricas

[30]:
```
# seleciona apenas as colunas numéricas
df_numerico = df_m_limpo.select_dtypes(include=[np.number])

# calcula a matriz de correlação
correlation_matrix = df_numerico.corr()

fig_heatmap = px.imshow(correlation_matrix,
                        title="Heatmap de Correlação entre Variáveis",
                        labels=dict(color="Correlação"),
                        color_continuous_scale='Viridis')

fig_heatmap.show()
```

1.4.2 Relação entre variáveis numéricas e categóricas

[14]:
```
# one hot enconding para transformar variáveis categóticas em numéricas
df_var_encoded = pd.get_dummies(df_m_limpo, drop_first=True)

df_var_encoded
```

[14]:

	ID_MOVIMENTO	CD_MOVIMENTO	ID_ESTACAO_ORIGEM	ID_ESTACAO_DESTINO	\
0	177.0	80.0	9.0	29.0	
1	179.0	81.0	9.0	29.0	
2	197.0	90.0	9.0	38.0	
3	199.0	91.0	9.0	38.0	
4	211.0	96.0	9.0	48.0	
…	…	…	…	…	
1219219	2267713.0	141785.0	91.0	23.0	
1219220	2267714.0	141786.0	91.0	24.0	
1219221	2267715.0	141787.0	91.0	24.0	
1219222	2268970.0	142456.0	91.0	79.0	
1219223	2268972.0	142457.0	91.0	79.0	

	CD_HORARIO_FK	CD_HORARIO_PARTIDA_FK	CD_HORARIO_CHEGADA_FK	\
0	39.0	39.0	39.0	
1	40.0	40.0	40.0	
2	39.0	39.0	41.0	
3	40.0	40.0	42.0	
4	39.0	39.0	43.0	

```
…                …                    …                    …
1219219          12.0                    12.0                    12.0
1219220          11.0                    11.0                    12.0
1219221          12.0                    12.0                    13.0
1219222          48.0                    48.0                    49.0
1219223          47.0                    47.0                    48.0

         NR_AJUSTE_MINUTO  NR_PASSAGEIRO  FL_SENTIDO  …  \
0                     8.0       1.432489         1.0  …
1                     8.0       1.432489         1.0  …
2                    28.0       1.023207         1.0  …
3                    28.0       1.023207         1.0  …
4                    56.0       1.432489         1.0  …
…                     …             …          …  …
1219219              14.0       0.420455         0.0  …
1219220              19.0       0.420455         0.0  …
1219221              19.0       0.420455         0.0  …
1219222              27.0       0.362745         0.0  …
1219223              27.0       0.181373         0.0  …

         TX_ESTACAO_Vila Olímpia  TX_ESTACAO_Várzea Paulista  \
0                          False                       False
1                          False                       False
2                          False                       False
3                          False                       False
4                          False                       False
…                          …                           …
1219219                    False                       False
1219220                    False                       False
1219221                    False                       False
1219222                    False                       False
1219223                    False                       False

         TX_ESTACAO_Água Branca  CD_TRECHO_A2  CD_TRECHO_B1  CD_TRECHO_C1  \
0                         False         False         False         False
1                         False         False         False         False
2                         False         False         False         False
3                         False         False         False         False
4                         False         False         False         False
…                         …             …             …             …
1219219                   False         False         False         False
1219220                   False         False         False         False
1219221                   False         False         False         False
1219222                   False         False         False         False
1219223                   False         False         False         False

         CD_TRECHO_D1  CD_TRECHO_E1  CD_TRECHO_E2  CD_TRECHO_F1
```

```
0              False         False         False         False
1              False         False         False         False
2              False         False         False         False
3              False         False         False         False
4              False         False         False         False
…                …             …             …             …
1219219        False         False         False         False
1219220        False         False         False         False
1219221        False         False         False         False
1219222        False         False         False         False
1219223        False         False         False         False

[1219224 rows x 487 columns]
```

[15]:
```python
# identifica colunas categóricas com apenas duas variáveis
colunas_binarias = [col for col in df_m_limpo.columns
                    if df_m_limpo[col].dtype == 'object'
                    and df_m_limpo[col].dropna().nunique() == 2]

print("Colunas categóricas com apenas duas variáveis:")
print(colunas_binarias)

# identifica colunas categóricas com mais de duas variáveis
colunas_multiclasse = [col for col in df_m_limpo.columns
                       if df_m_limpo[col].dtype == 'object'
                       and df_m_limpo[col].dropna().nunique() > 2]

print("\nColunas categóricas com mais de duas variações variáveis:")
print(colunas_multiclasse)
```

```
Colunas categóricas com apenas duas variáveis:
[]

Colunas categóricas com mais de duas variações variáveis:
['TX_ESTACAO_ORIGEM', 'TX_ESTACAO_DESTINO', 'ID_TRECHO_ORIGEM',
'ID_TRECHO_DESTINO', 'TX_DIAS', 'TX_PREFIXO', 'TX_ESTACAO', 'CD_TRECHO']
```

1.4.3 Análise Univariada

NR_AJUSTE_MINUTO

[12]:
```python
# Histograma para uma variável numérica
fig = px.histogram(df_m_limpo, x='NR_AJUSTE_MINUTO')
fig.show()

# Box plot para uma variável numérica
fig = px.box(df_m_limpo, y='NR_AJUSTE_MINUTO')
fig.show()
```

O significado exato do valor de `NR_AJUSTE_MINUTO` ainda é desconhecido. Para correnaciona-lo com a análise de ocorrências e falhas, essa váriavel pode representar ajustes durante a operação, como mudanças de tempo devido a manutenção (referente à alguma falha) na linha ou controle de fluxo de trens em determinadas seções da rota.

Como observado no box plot acima, o valor ´65´ é um outlier, porém se analizarmos o histograma gerado anteriormente podemos identificar um número relativamente alto para esse valor, o que pode ser um ponto de atenção.

1.4.4 Análise Bivariada

TX_ESTACAO_ORIGEM E TX_ESTACAO_DESTINO

[13]:
```
fig_bivariada = px.scatter(df_m_limpo, x='TX_ESTACAO_ORIGEM',␣
 ↪y='TX_ESTACAO_DESTINO',
                           title='Análise Bivariada entre X e Y')
fig_bivariada.show()
```

O gráfio acima mostra a relação entre trechos de trens com base nos pontos de origem e destino, mostrando quais as estações com maior movimento entre si.

1.4.5 Análise de Outliers

[18]:
```
# visualiza a quantidade de valores ausentes em cada coluna
valores_ausentes = df_m_limpo.isnull().sum()

fig_valores_ausentes = px.bar(valores_ausentes,
                              title="Valores Ausentes por Coluna",
                              labels={"index": "Coluna", "value": "Número de␣
 ↪Valores Ausentes"})
fig_valores_ausentes.show()
```

Grande número de colunas possuem valores nulos.

[7]:
```
# selecionar apenas as colunas numéricas do dataframe
df_numerico = df_m_limpo.select_dtypes(include=['number'])

Q1 = df_numerico.quantile(0.25)
Q3 = df_numerico.quantile(0.75)
IQR = Q3 - Q1

outliers = (df_numerico < (Q1 - 1.5 * IQR)) | (df_numerico > (Q3 + 1.5 * IQR))

# visualiza o número de outliers por coluna
fig_outliers = px.bar(outliers.sum(),
                      title="Número de Outliers por Coluna",
                      labels={"index": "Coluna", "value": "Número de Outliers"})
fig_outliers.show()
```

Análise da coluna `NR_AJUSTE_MINUTO` feita acima

[17]:
```
# elimina os outliers identificados
df_sem_outliers = df_numerico[~((df_numerico < (Q1 - 1.5 * IQR)) | (df_numerico␣
 ↪> (Q3 + 1.5 * IQR))).any(axis=1)]

df_sem_outliers
```

[17]:

	ID_MOVIMENTO	CD_MOVIMENTO	ID_ESTACAO_ORIGEM	ID_ESTACAO_DESTINO	\
0	177.0	80.0	9.0	29.0	
1	179.0	81.0	9.0	29.0	
2	197.0	90.0	9.0	38.0	
3	199.0	91.0	9.0	38.0	
4	211.0	96.0	9.0	48.0	
…	…	…	…	…	
1219210	2267701.0	141770.0	91.0	79.0	
1219213	2267705.0	141773.0	91.0	79.0	
1219214	2267707.0	141774.0	91.0	79.0	
1219222	2268970.0	142456.0	91.0	79.0	
1219223	2268972.0	142457.0	91.0	79.0	

	CD_HORARIO_FK	CD_HORARIO_PARTIDA_FK	CD_HORARIO_CHEGADA_FK	\
0	39.0	39.0	39.0	
1	40.0	40.0	40.0	
2	39.0	39.0	41.0	
3	40.0	40.0	42.0	
4	39.0	39.0	43.0	
…	…	…	…	
1219210	10.0	10.0	11.0	
1219213	9.0	9.0	10.0	
1219214	10.0	10.0	11.0	
1219222	48.0	48.0	49.0	
1219223	47.0	47.0	48.0	

	NR_AJUSTE_MINUTO	NR_PASSAGEIRO	FL_SENTIDO	…	TEMPO_PERC	\
0	8.0	1.432489	1.0	…	NaN	
1	8.0	1.432489	1.0	…	NaN	
2	28.0	1.023207	1.0	…	NaN	
3	28.0	1.023207	1.0	…	NaN	
4	56.0	1.432489	1.0	…	NaN	
…	…	…	…	…	…	
1219210	27.0	0.387500	0.0	…	NaN	
1219213	27.0	0.387500	0.0	…	NaN	
1219214	27.0	0.387500	0.0	…	NaN	
1219222	27.0	0.362745	0.0	…	NaN	
1219223	27.0	0.181373	0.0	…	NaN	

	TEMPO_TRANSF	FL_ESTACAO	ID_ESTACAO	ID_LINHA	NR_SEQ_ESTACAOLINHA	\
0	NaN	NaN	NaN	NaN	NaN	

```
1                NaN          NaN          NaN       NaN                NaN
2                NaN          NaN          NaN       NaN                NaN
3                NaN          NaN          NaN       NaN                NaN
4                NaN          NaN          NaN       NaN                NaN
…                …            …            …         …                  …
1219210          NaN          NaN          NaN       NaN                NaN
1219213          NaN          NaN          NaN       NaN                NaN
1219214          NaN          NaN          NaN       NaN                NaN
1219222          NaN          NaN          NaN       NaN                NaN
1219223          NaN          NaN          NaN       NaN                NaN

         ID_REGISTRO  ID_CODUSUARIO  FL_FIM_TRECHO  ID_LINHA_FK
0                NaN            NaN            NaN          NaN
1                NaN            NaN            NaN          NaN
2                NaN            NaN            NaN          NaN
3                NaN            NaN            NaN          NaN
4                NaN            NaN            NaN          NaN
…                …              …              …            …
1219210          NaN            NaN            NaN          NaN
1219213          NaN            NaN            NaN          NaN
1219214          NaN            NaN            NaN          NaN
1219222          NaN            NaN            NaN          NaN
1219223          NaN            NaN            NaN          NaN

[899809 rows x 28 columns]
```

[19]:
```python
# Histograma para uma variável numérica
fig = px.histogram(df_sem_outliers, x='NR_AJUSTE_MINUTO')
fig.show()

# Box plot para uma variável numérica
fig = px.box(df_sem_outliers, y='NR_AJUSTE_MINUTO')
fig.show()
```

Os gráficos de análise da coluna ***NR_AJUSTE_MINUTO*** *foram refeitos após a remoção dos outliers previamente identificados. Na primeira versão dos gráficos, o valor '65' era claramente visível como um outlier. No entanto, ao gerar os gráficos com um novo dataframe sem os outliers, as ocorrências do valor '62' também foram eliminadas. Ao observar o primeiro histograma, nota-se que o valor '62' possuía poucas ocorrências em comparação ao total. Isso pode indicar que esses valores estavam distorcendo a distribuição dos dados, e a sua remoção tornou a análise mais precisa, destacando melhor o comportamento da variável.*

1.5 Visualização com Gráficos

[14]:
```python
df_freq = df_m_limpo.dropna(subset=['TX_ESTACAO_ORIGEM', 'TX_ESTACAO_DESTINO'])

# frequência de movimentos por estação de origem
```

```
freq_origem = df_freq['TX_ESTACAO_ORIGEM'].value_counts().reset_index()
freq_origem.columns = ['Estação de Origem', 'Frequência']

fig_origem = px.bar(freq_origem, x='Estação de Origem', y='Frequência',
                    title='Frequência de Movimentos por Estação de Origem',
                    color='Frequência', text='Frequência')

fig_origem.update_traces(texttemplate='%{text}', textposition='outside')
fig_origem.show()

# frequência de movimentos por estação de destino
freq_destino = df_freq['TX_ESTACAO_DESTINO'].value_counts().reset_index()
freq_destino.columns = ['Estação de Destino', 'Frequência']

fig_destino = px.bar(freq_destino, x='Estação de Destino', y='Frequência',
                     title='Frequência de Movimentos por Estação de Destino',
                     color='Frequência', text='Frequência')

fig_destino.update_traces(texttemplate='%{text}', textposition='outside')
fig_destino.show()
```

Hipótese: As estações com maior frequência de movimentos podem indicar maior demanda ou importância, significando também que comportam um maior volume de passageiros diariamente.

Análise de Ocorrências e Falhas: Nesse caso, analisar se as estações com maior movimento têm uma taxa elevada de ocorrências ou falhas para direcionar esforços às estações mais frequentadas.

[16]:
```
df_dias = df_m_limpo.dropna(subset=['TX_DIAS'])

# frequência de movimentos por dia
freq_dias = df_dias['TX_DIAS'].value_counts().reset_index()
freq_dias.columns = ['Dia', 'Frequência']

fig_dias = px.bar(freq_dias, x='Dia', y='Frequência',
                  title='Frequência de Movimentos por Dia',
                  color='Frequência', text='Frequência')

fig_dias.update_traces(texttemplate='%{text}', textposition='outside')
fig_dias.show()
```

Hipótese: A frequência de movimentos pode variar com os dias da semana, indicando padrões de maior uso do serviço de trens. O que mostra quais dias há um maior volume de passageiros em circulação.

Análise de Ocorrências e Falhas: Assim, examinar se os dias com maior movimento também apresentam um aumento nas falhas pode indicar a necessidade de mais recursos ou melhorias operacionais, além de realocação de esforços para tais dias, evitando problemas futuros com falta de disponibilidade ou assistência em tal dia.

1.6 Conclusões

A tabela analisada, embora não esteja diretamente relacionada ao segmento de *ocorrências e falhas*, trouxe insights valiosos que podem ser úteis no tratamento de dados voltados para esse tema. Por exemplo, as análises de frequência de movimento por estação e por dia da semana podem servir como pontos de atenção, permitindo a alocação estratégica de equipes de assistência e destacando trechos ou dias com maior risco de ocorrência. Esses dados foram inicialmente analisados de forma isolada, mas, se cruzados com uma tabela específica de ocorrências e falhas, é possível obter insights mais concretos e relevantes sobre essas situações.

1.6.1 Convertendo para Parquet

[21]:
```
import pyarrow.parquet as pq

df_m_limpo.to_parquet('CPA.parquet', index=False)
```