Data Exploration

This section will be quite short, since we have explored most of the data in the dataCleaning.ipynb notebook file.

Here we will try and look at the amount of categories, and how the resumes are spread between these categories.

In [31]:
```
import pandas as pd
import numpy as np
import pickle
```

In this file we will explore the already cleaned data, which we load with pickle

In [32]:
```
# Loading the data
with open('Data/Dataframes/newDF.pkl', 'rb') as f:
    df = pickle.load(f)
```

We will quickly look at the spread of the data in the columns. As we are working with resumes as text, we cant get many interesting visualizations yet.

In [33]:
```
from matplotlib import pyplot as plt
import seaborn as sns
```

In [34]:
```
# Lets visualize the categories of resumes using the one hot encoding
print(df.columns)
```

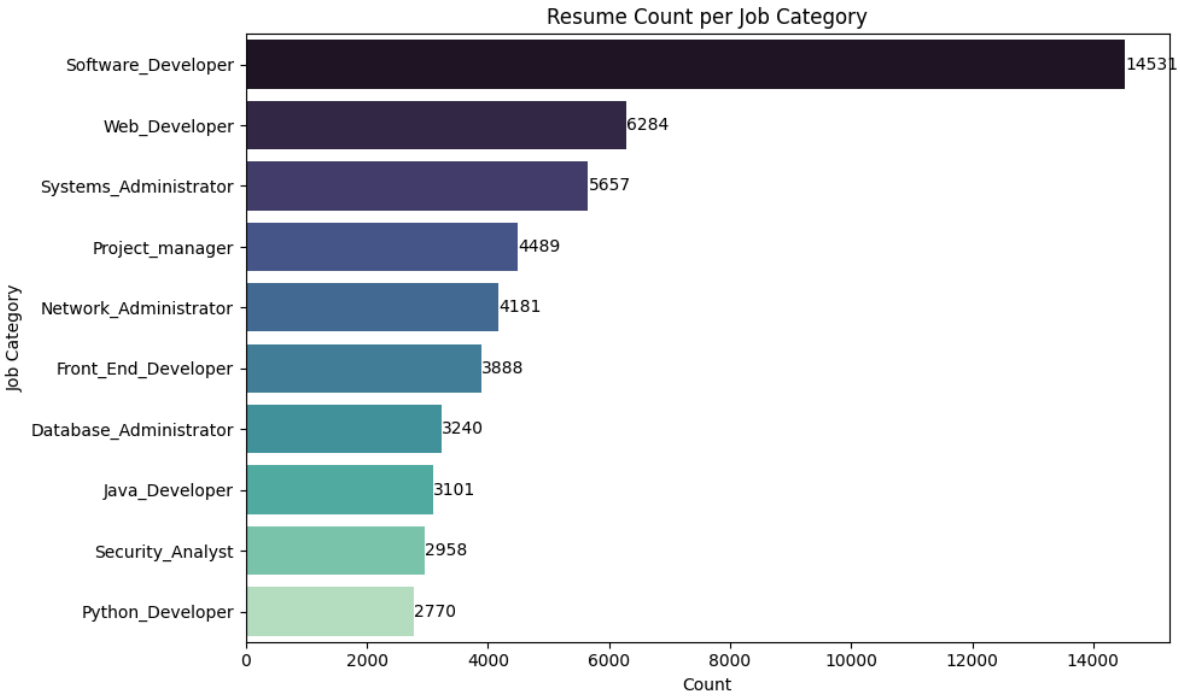
```
Index(['ID', 'Resume', 'Label', 'Software_Developer', 'Database_Administrator',
       'Systems_Administrator', 'Project_manager', 'Web_Developer',
       'Network_Administrator', 'Security_Analyst', 'Python_Developer',
       'Java_Developer', 'Front_End_Developer', 'TextLen'],
      dtype='object')
```

In [35]:
```
category_columns = [
    'Software_Developer', 'Database_Administrator', 'Systems_Administrator',
    'Project_manager', 'Web_Developer', 'Network_Administrator',
    'Security_Analyst', 'Python_Developer', 'Java_Developer',
    'Front_End_Developer'
]
```

In [36]:
```
category_counts = df[category_columns].sum().sort_values(ascending=False)
```

In [37]:
```
plt.figure(figsize=(10, 6))
sns.barplot(x=category_counts.values, y=category_counts.index, palette='mako')
for i, v in enumerate(category_counts.values):
    plt.text(v + 0.5, i, str(v), color='black', va='center')
plt.title("Resume Count per Job Category")
plt.xlabel("Count")
plt.ylabel("Job Category")
plt.tight_layout()
plt.show()
```

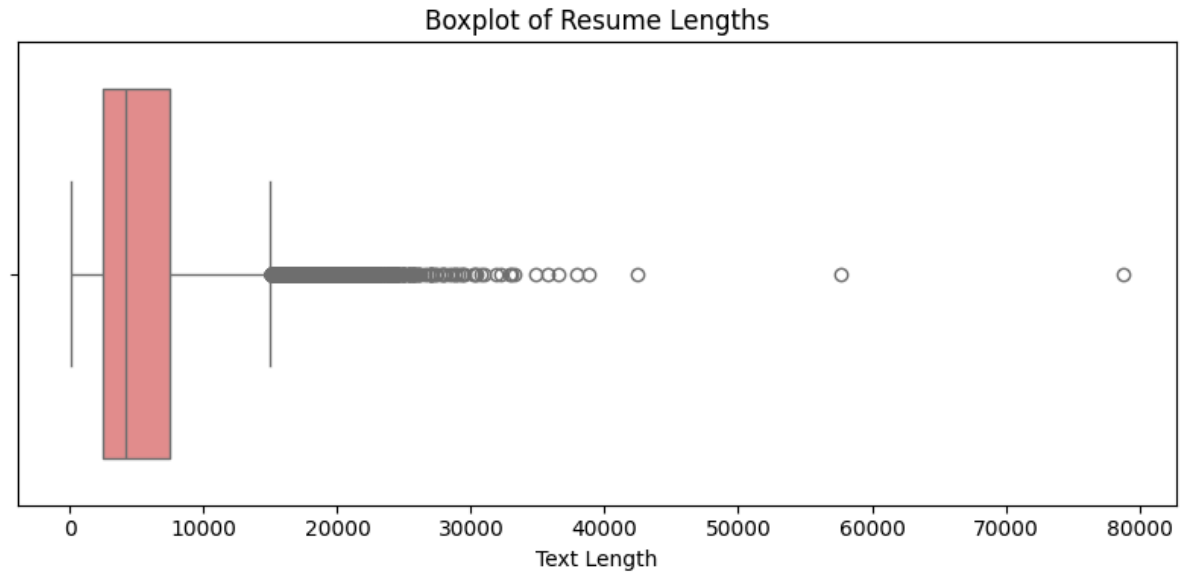
```
C:\Users\pelle\AppData\Local\Temp\ipykernel_4792\1369709213.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.
14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=category_counts.values, y=category_counts.index, palette='mako')
```



From this graph we can tell that there is a big discrepancy between the different resume job categories. This can maybe affect the training of the models later on, but we dont want to change more of the dataframe.

Especially since it makes sense, that there will be more job applications with software devs compared to python devs.

In [38]:
```
plt.figure(figsize=(8, 4))
sns.boxplot(x=df['TextLen'], color='lightcoral')
plt.title("Boxplot of Resume Lengths")
plt.xlabel("Text Length")
plt.tight_layout()
plt.show()
```



Our spread of resume lengths look fine. There are some outliers, but we are not too worried with those, since we will embed the texts later on when training the models.