In [ ]:

```python
import os
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
import warnings

sns.set_theme(style="ticks")
warnings.simplefilter(action='ignore', category=FutureWarning)
```

1) Data ingestion and preparation

Steps:

- Reading data from csv into pandas dataframe
- Checking for missing values in all columns
- Converte datetime string to date
- Correct misspelled "CustomerName"

In [ ]:

```python
# Reading input data
data_path = os.path.abspath(os.path.join(os.path.abspath(""), '..', 'data'))
df = pd.read_csv(f'{data_path}/usageData.csv')
#display(df)
```

In [ ]:

```python
# Check if we have any missing data
print(df.isna().sum())
```

```
Date              0
Region            0
VMSeries          0
unit              0
ComputeUsage      0
SubscriptionID    0
CustomerName      0
dtype: int64
```

In [ ]:

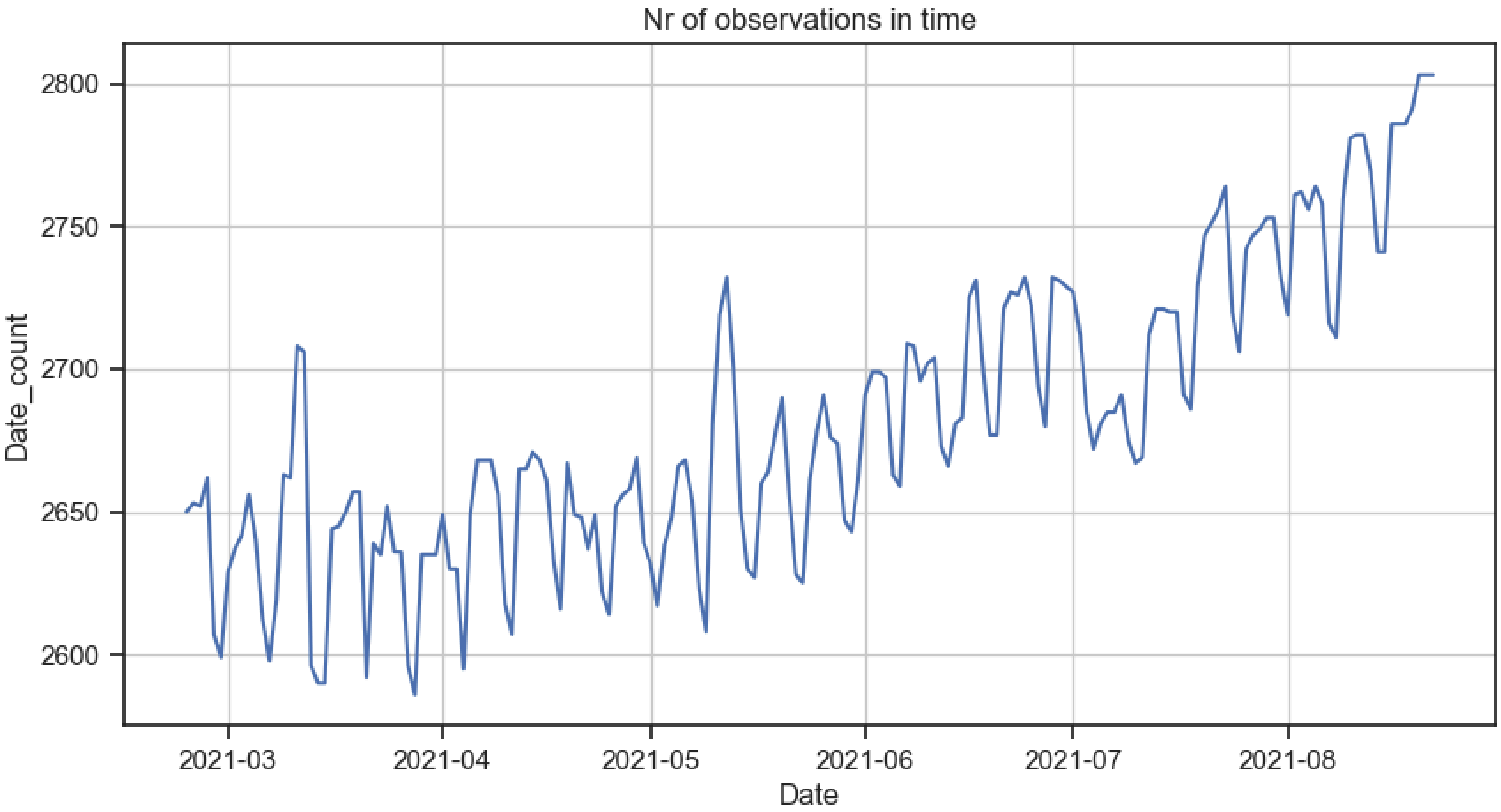
```python
# Data preparation
df['Date'] = pd.to_datetime(df['Date']).dt.date
```

```
df['CustomerName'] = df['CustomerName'].str.replace('CustomberB','CustomerB')
```

2) Visialization through time

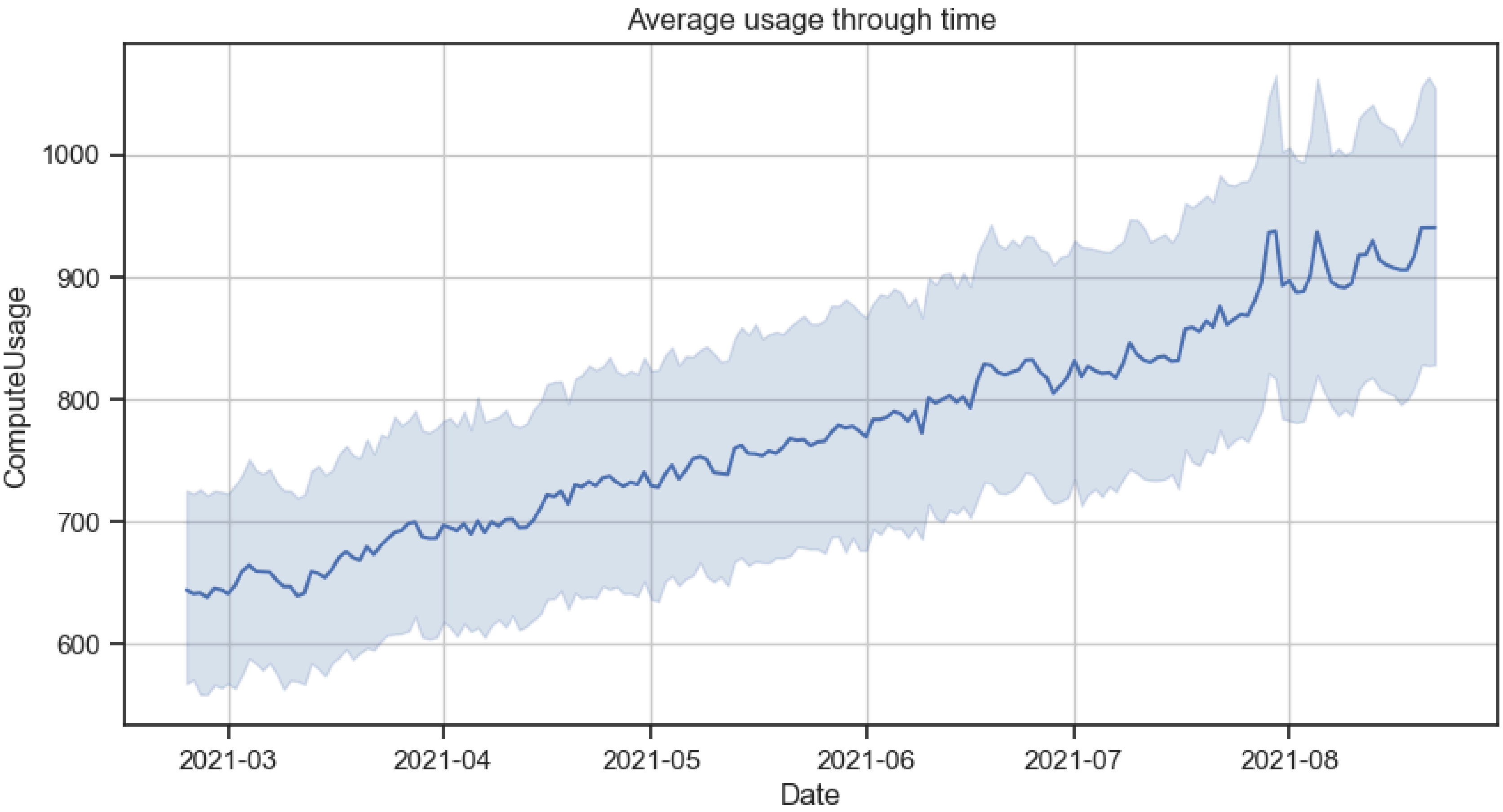
In [ ]:

```
# Nr of observations through time
df_plt = df.groupby('Date').agg(Date_count=pd.NamedAgg(column="Date", aggfunc="count"),
                                ComputeUsage_avg=pd.NamedAgg(column="ComputeUsage", aggfunc="mean"),
                                ComputeUsage_min=pd.NamedAgg(column="ComputeUsage", aggfunc="min"),
                                ComputeUsage_max=pd.NamedAgg(column="ComputeUsage", aggfunc="max")
                                ) \
                           .reset_index()

plt.figure(figsize=(10,5))
sns.lineplot(data=df_plt, x='Date', y='Date_count')
plt.grid()
plt.title("Nr of observations in time")
plt.show()
```

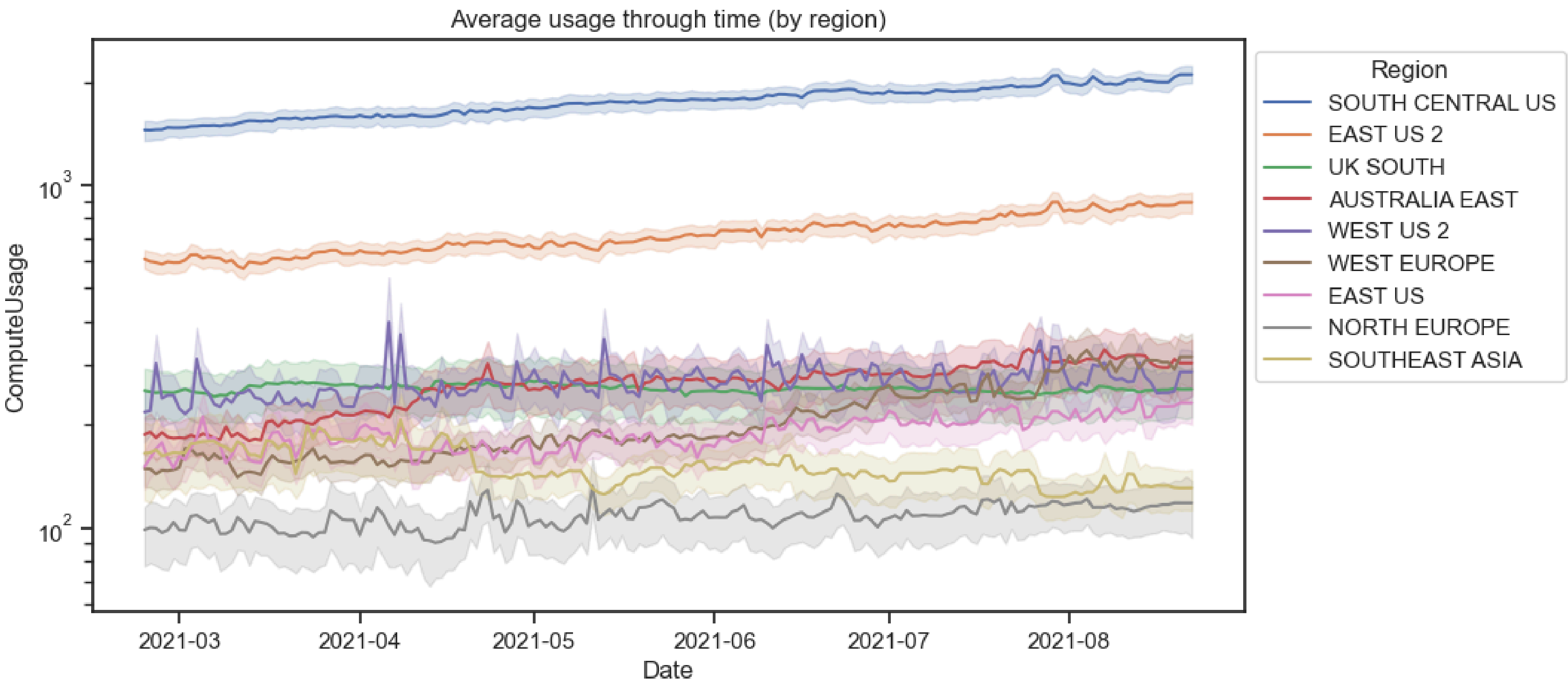


- The number of observations are incresing with time

```
In [ ]: plt.figure(figsize=(10,5))
        sns.lineplot(data=df, x='Date', y='ComputeUsage', estimator='mean', errorbar=('ci', 95))
        plt.grid()
        plt.title("Average usage through time")
        plt.show()
```

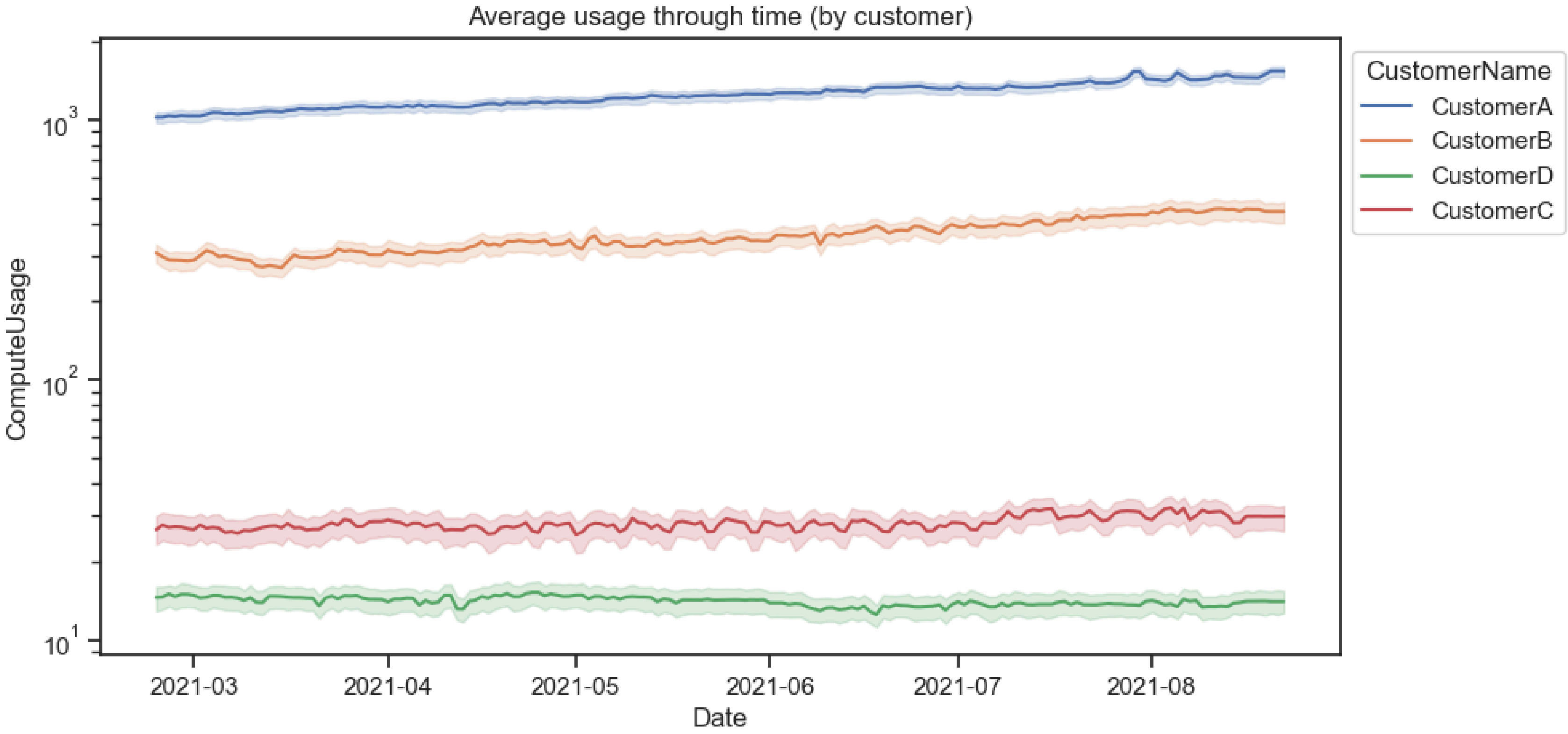


```
In [ ]: plt.figure(figsize=(10,5))
        ax = sns.lineplot(data=df, x='Date', y='ComputeUsage', estimator='mean', errorbar=('ci', 50), hue='Region')
        sns.move_legend(ax, "upper left", bbox_to_anchor=(1, 1))
        ax.set(yscale='log')
        plt.title("Average usage through time (by region)")
        plt.show()
```



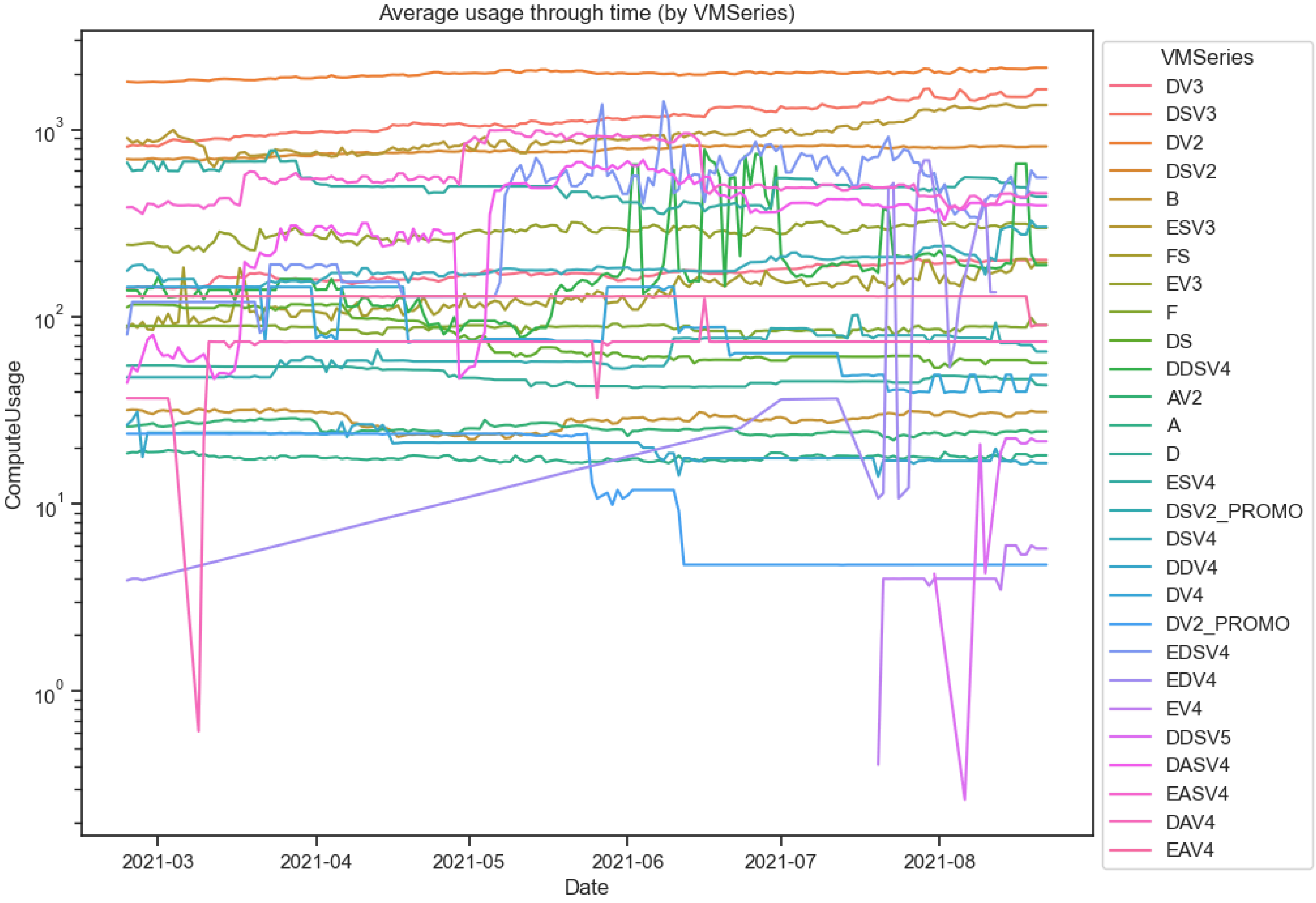
In [ ]:

```
plt.figure(figsize=(10,5))
ax = sns.lineplot(data=df, x='Date', y='ComputeUsage', estimator='mean', errorbar=('ci', 50), hue='CustomerName')
sns.move_legend(ax, "upper left", bbox_to_anchor=(1, 1))
ax.set(yscale='log')
plt.title("Average usage through time (by customer)")
plt.show()
```



In [ ]:

```
plt.figure(figsize=(10,8))
ax = sns.lineplot(data=df, x='Date', y='ComputeUsage', estimator='mean', errorbar=None, hue='VMSeries')
sns.move_legend(ax, "upper left", bbox_to_anchor=(1, 1))
ax.set(yscale='log')
plt.title("Average usage through time (by VMSeries)")
plt.show()
```



3) Visualization by customer

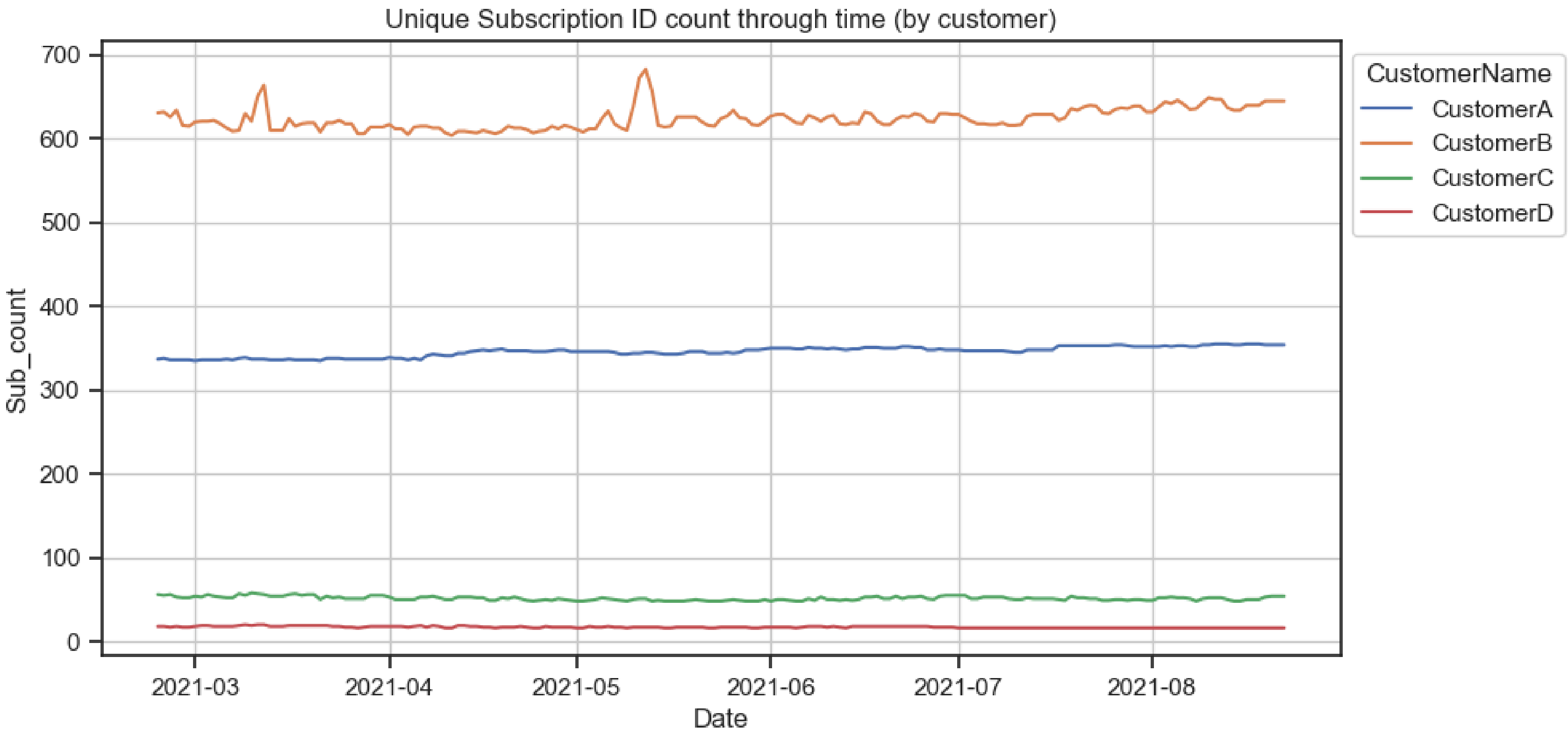
In [ ]:

```
# Number of subscription IDs by customer
display(df.groupby('CustomerName')['SubscriptionID'].nunique())
```

```
CustomerName
CustomerA    393
CustomerB    902
CustomerC     74
CustomerD     22
Name: SubscriptionID, dtype: int64
```

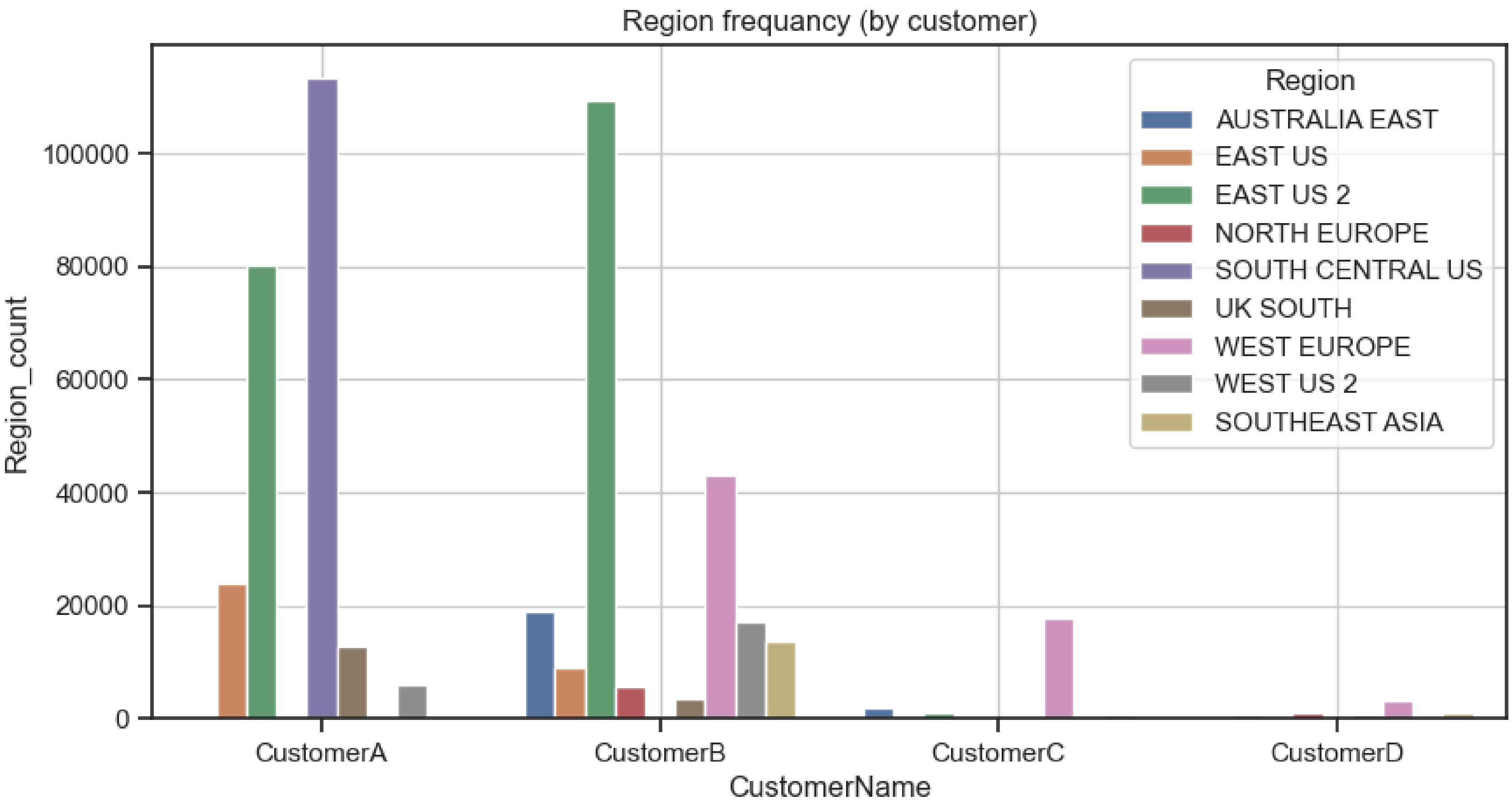
In [ ]:

```
df_plt = df.groupby(['CustomerName', 'Date']).agg(Sub_count=pd.NamedAgg(column="SubscriptionID", aggfunc="nunique"))

plt.figure(figsize=(10,5))
ax = sns.lineplot(data=df_plt, x='Date', y='Sub_count',  hue='CustomerName')
sns.move_legend(ax, "upper left", bbox_to_anchor=(1, 1))
plt.grid()
plt.title("Unique Subscription ID count through time (by customer)")
plt.show()
```



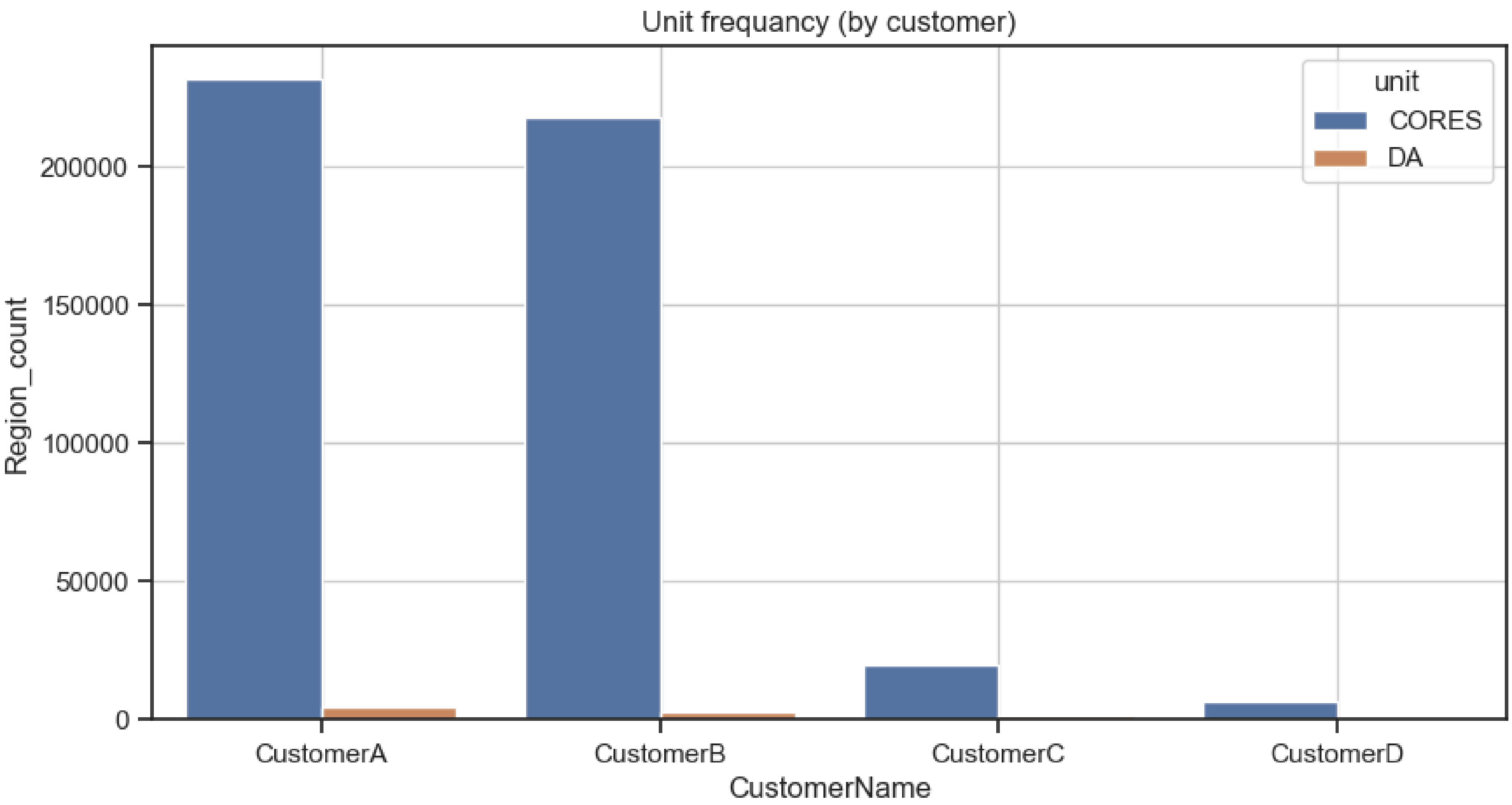
In [ ]:

```
df_plt = df.groupby(['CustomerName', 'Region']).agg(Region_count=pd.NamedAgg(column="Region", aggfunc="count")).reset

plt.figure(figsize=(10,5))
sns.barplot(data = df_plt, x="CustomerName", y="Region_count", hue="Region")
plt.grid()
plt.title("Region frequancy (by customer)")
plt.show()
```



In [ ]:

```
df_plt = df.groupby(['CustomerName', 'unit']).agg(Region_count=pd.NamedAgg(column="unit", aggfunc="count")).reset_in

plt.figure(figsize=(10,5))
sns.barplot(data = df_plt, x="CustomerName", y="Region_count", hue="unit")
plt.grid()
plt.title("Unit frequancy (by customer)")
plt.show()
```



4) Conclusions

- The number of observations in the dataset is increasing with time
- Average ComputeUsage has a positive trend
- Most of the ComputeUsage comes from "South Central US" and "East US2" regions
- Customers A, B and C are the customers with the higest usage
- Customers A and B also have the highest number of subscriptions
- Customer A has half the number of subscriptions of customer B, but has significantly higher ComputeSusage

- Customers A and B use mostly resources from US region (South Central and East 2)
- Customers C and D use mostly resources from West Europe region