SalesPlaybookDS5640: GitHub Data Loader & Cleaning Starter

This document contains a shared function that allows each team member to load the project datasets directly from the GitHub repository. This helps avoid issues with version control and streamlines collaboration. Below you'll also find starter code for working with each dataset.

Note: The sample cleaning code includes a line that drops rows where all values are missing. If you'd prefer to keep these rows for your dataset, feel free to remove that line of code.

Universal GitHub Loader Function

```
import pandas as pd

def load_csv_from_github(file_name):
    base_url = "https://raw.githubusercontent.com/marymorkos/SalesPlaybookDS5640/refs/heads/main/"
    return pd.read_csv(base_url + file_name)
```

Deals Dataset

```
deals_df = load_csv_from_github("anonymized_hubspot_deals.csv")
deals_df.head()

# Example cleaning steps
deals_df.columns = [col.strip().lower().replace(" ", "_") for col in deals_df.columns]
deals_df.dropna(how="all", inplace=True)  # Optional: remove this line if you want to keep fully empty rows
```

Companies Dataset

```
companies_df = load_csv_from_github("anonymized_hubspot_companies.csv")
companies_df.head()

# Example cleaning steps
companies_df.columns = [col.strip().lower().replace(" ", "_") for col in companies_df.columns]
companies_df.dropna(how="all", inplace=True)  # Optional: remove this line if you want to keep fully empty rows
```

Tickets Dataset

```
tickets_df = load_csv_from_github("anonymized_hubspot_tickets.csv")
tickets_df.head()

# Example cleaning steps
tickets_df.columns = [col.strip().lower().replace(" ", "_") for col in tickets_df.columns]
tickets_df.dropna(how="all", inplace=True)  # Optional: remove this line if you want to keep fully empty rows
```