Assignment 03

Norah Jones

November 21, 2024

1 Load the Dataset

The instruction below provides you with general keywords for columns used in the lightcast file. See the data schema generated after the load dataset code above to use proper column name. For each visualization, **customize colors, fonts, and styles** to avoid a **2.5-point deduction**. Also, **provide a two-sentence explanation** describing key insights drawn from the graph.

1. **Load the Raw Dataset**:
    - Use Pyspark to the `lightcast_data.csv` file into a DataFrame:
    - You can reuse the previous code.
    - Copying code from your friend constitutes plagiarism. DO NOT DO THIS.

```
from pyspark.sql import SparkSession
import pandas as pd
import plotly.express as px
import plotly.io as pio
import numpy as np

np.random.seed(42)

pio.renderers.default = "notebook"

# Initialize Spark Session
spark = SparkSession.builder.appName("LightcastData").getOrCreate()

# Load Data
df = spark.read.option("header", "true").option("inferSchema", "true").option("multiLine","t

# Show Schema and Sample Data
```

```
# print("---This is Diagnostic check, No need to print it in the final doc---")

# df.printSchema() # comment this line when rendering the submission
# df.show(5)
```

```
[Stage 8:>                                                          (0 + 1) / 1]
```

2 Data Preparation

3 Salary Distribution by Industry and Employment Type

- Compare salary variations across industries.
- **Filter the dataset**
  - Remove records where **salary is missing or zero**.
- **Aggregate Data**
  - Group by **NAICS industry codes**.
  - Group by **employment type** and compute salary distribution.
  - Calculate **salary percentiles** (25th, 50th, 75th) for each group.
- **Visualize results**
  - Create a **box plot** where:
    - **X-axis** = `NAICS2_NAME`
    - **Y-axis** = `SALARY_FROM`, or `SALARY_TO`, or `SALARY`
    - Group by `EMPLOYMENT_TYPE_NAME`.
  - Customize colors, fonts, and styles.
- **Explanation:** Write two sentences about what the graph reveals.

4 Salary Analysis by ONET Occupation Type (Bubble Chart)

- Analyze how salaries differ across ONET occupation types.
- **Aggregate Data**
  - Compute **median salary** for each occupation in the **ONET taxonomy**.
- **Visualize results**
  - Create a **bubble chart** where:

    - **X-axis** = `ONET_NAME`
    - **Y-axis** = `Median Salary`
    - **Size** = Number of job postings
  - Apply custom colors and font styles.
- **Explanation:** Write two sentences about what the graph reveals.

5 Salary by Education Level

- Create two groups:
  - **Bachelor's or lower** (Bachelor's, GED, Associate, No Education Listed)
  - **Master's or PhD** (Master's degree, Ph.D. or professional degree)
- Plot scatter plots for each group using, `MAX_YEARS_EXPERIENCE` (with jitter), `Average_Salary`, `LOT_V6_SPECIALIZED_OCCUPATION_NAME`
- Then, plot histograms overlaid with KDE curves for each group.
- This would generate two scatter plots and two histograms.
- **After each graph, add a short explanation** of key insights.

6 Salary by Remote Work Type

- Split into three groups based on `REMOTE_TYPE_NAME`:
  - Remote
  - Hybrid
  - Onsite (includes `[None]` and blank)
- Plot scatter plots for each group using, `MAX_YEARS_EXPERIENCE` (with jitter), `Average_Salary`, `LOT_V6_SPECIALIZED_OCCUPATION_NAME`
- Also, create salary histograms for all three groups.
- **After each graph, briefly describe any patterns or comparisons.**