Database Design and Implementation Project

Introduction:

In this report, we detail the process of reproducing the approach and evaluation methodology presented in the paper titled “SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics”. The paper outlines an algorithm based on Shared-based Optimization (through query rewriting) and Pruning-based Optimization (using Hoeffding-Serfling inequality) for optimizing aggregate queries, with the evaluation conducted using the census dataset.

Dataset:

We used the UCI Machine Learning Repository's census dataset, which contains demographic information about individuals such as age, education, marital status, occupation, etc. This dataset comprises 32,561 entries and 15 features, including both numerical and categorical variables.

<u>Numerical variables:</u> Age, FinalWeight, Education_Num , Capital Gain, Capital Loss, Hours_per_week.
<u>Categorical variables:</u> Income, WorkClass, Education, Occupation, Relationship, Race, Native_Country, Sex.

Data Preprocessing:

Handling Missing Values :

- Before applying the optimization algorithms, we preprocessed the dataset to handle missing values.
- We imputed the missing values with mode as it doesn't disturb the probability distribution much. We have done the mode imputation on the columns which had the missing values.

Marital Status Mapping:

- Mapping the marital status into two categories: 'Married' and 'Unmarried'.
- We are categorizing individuals as "Unmarried" only if their marital status is "Never-married", while all other marital statuses are considered as "Married".
- This decision is made because individuals classified as "Never-married" are more likely to have never been married, ensuring a clear distinction between married and unmarried groups.
- This approach aims to minimize data loss and ensure that the analysis accurately captures individuals who are currently married and unmarried without overlooking any potential variations in marital status over time.

Algorithm Implementation:

We implemented the Shared-based Optimization and Pruning-based Optimization algorithms described in Sections 4.1 and 4.2 of the paper, respectively. The algorithms aim to find the top-5 aggregate views by utility measure (K-L Divergence) using user-specified queries for married and unmarried people.

- ***Step 1: Database Connection:***
    - We establish a connection to an SQLite database name 'census_data.db', which contains the census dataset that is pre processed.

- ***Step 2: SQL Query Construction (Shared-based Optimization):***

○ We leverage shared-based optimization by constructing SQL queries that efficiently aggregate data for all numerical attributes grouped on sex and marital_status.

○ We select the specified categorical attributes and calculate aggregate functions (mean, sum, count, min, max) for all numeric attributes.

○ By grouping data based on all attributes, we ensure that the query is optimized for retrieving relevant information related to married individuals.

○ Similarly, for unmarried individuals, we construct a separate query using the same optimization principles.

Features Discarded: We have discarded the 'Relationship' feature as 'Relationship' and 'marital_status' have high correlation, as some values for relationships are exclusive to married or unmarried people. This might lead to high kl- divergence and show up in the top 5 plots, but logically speaking these plots are not very interesting as we already know that all the husbands and wives are married(as an example).

- ***Step 3: Query Execution and Data Retrieval:***

○ We execute the constructed SQL queries to retrieve the aggregated data for married and unmarried individuals from the database.

○ The fetched results are stored in separate variables, ensuring efficient data retrieval tailored to the marital status of individuals.

○ This approach optimizes query execution by targeting specific subsets of data based on marital status, enhancing performance and resource utilization.

Step 4: Data Processing:

○ Upon retrieval, we convert the aggregated data into pandas DataFrames for easier manipulation and analysis.

○ The DataFrames contain columns representing the specified categorical attributes and aggregate functions for numeric attributes, facilitating further processing and exploration.

○ This processing step ensures that the data is structured and organized for efficient analysis and interpretation.

Step 5: K-L Divergence Calculation (Pruning-based Optimization):

- We apply pruning-based optimization to identify relevant attributes and aggregate functions that exhibit significant differences between married and unmarried individuals.
- By calculating the K-L Divergence between probability distributions of common columns, we identify attributes with divergent distributions.
- We normalize the data into probability distributions and handle outliers by clipping values within a specified range, ensuring robust divergence calculations.
- The pruning condition is applied to consider only those columns with K-L Divergence scores greater than a certain threshold, focusing on the most relevant and informative attributes.

- ***Step 6: Result Interpretation:***
  - The K-L Divergence scores obtained through pruning-based optimization provide insights into the distinct patterns and distributions within the census dataset based on marital status.
  - By interpreting these scores, we identify attributes and aggregate functions that signify significant differences between married and unmarried individuals.
  - These insights contribute to a better understanding of demographic characteristics and their associations with marital status, facilitating informed decision-making and analysis.
  - The K-L Divergence score indicates the divergence between the probability distributions of attributes for married and unmarried individuals.
  - We apply a pruning condition to consider only those columns with K-L Divergence scores greater than a certain threshold.

Overall KL_Divergence for all over the threshold of kl_score of 0.000001:

```python
# Display the DataFrame
df_views
```

Out[21]:

	key	Attribute	KL Divergence
0	Occupation	mean_capital_loss	9.137374e+00
1	Education	max_capital_loss	2.516445e+00
2	Education	mean_capital_loss	2.486137e+00
3	Education	max_capital_gain	1.501893e+00
4	Education	mean_capital_gain	1.460488e+00
...	...	...	...
151	Sex	mean_Final_Weight	1.818393e-04
152	Income	mean_education_num	1.813080e-04
153	Sex	max_Final_Weight	7.388642e-05
154	Sex	mean_age	3.370779e-06
155	Income	mean_Final_Weight	2.333545e-07

156 rows × 3 columns

Top 5 Results:

After implementing the algorithms, we obtained the top-5 aggregate views for both married and unmarried individuals based on K-L Divergence. The views include various combinations of categorical_attributes such as income, work class, education, occupation, etc., with corresponding aggregate functions like mean, sum, count, min, and max for numerical attributes.

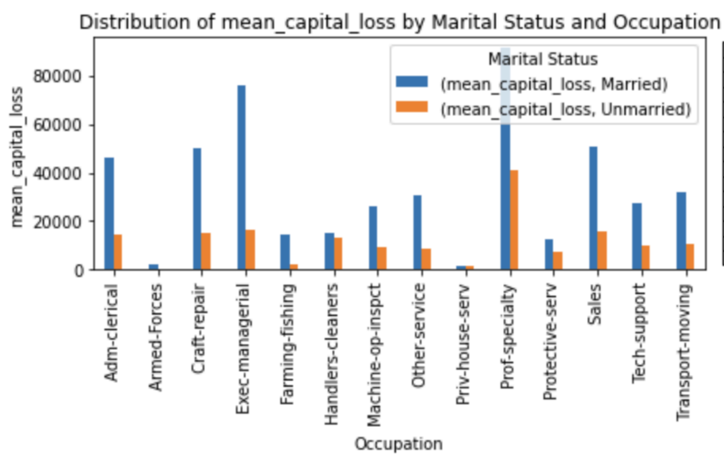
```python
# Display the DataFrame
df_top_views
```

Out[22]:

	key	Attribute	KL Divergence
0	Occupation	mean_capital_loss	9.137374
1	Education	max_capital_loss	2.516445
2	Education	mean_capital_loss	2.486137
3	Education	max_capital_gain	1.501893
4	Education	mean_capital_gain	1.460488

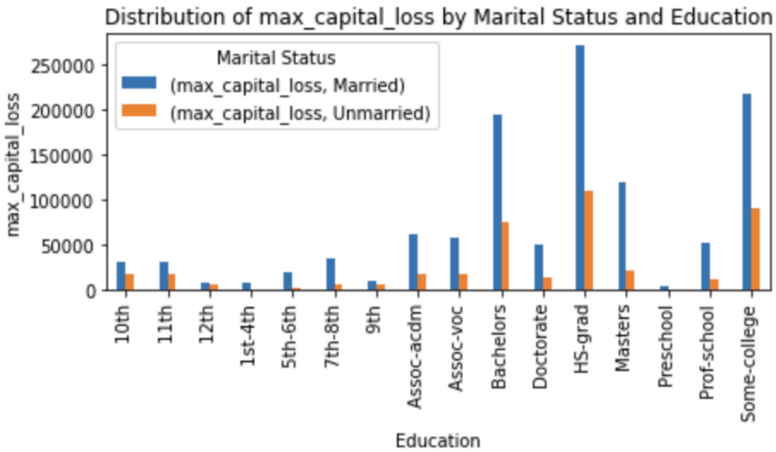
Results and Interpretation:

The K-L Divergence provides insights into the divergence between the probability distributions of attributes for married and unmarried individuals, helping identify relevant features for each group.
The optimized aggregate views, grouped by sex, highlight distinct patterns and distributions within the dataset for each marital status category.
Lets see briefly about each of them, listed in descending order based on their KL-Divergence values:



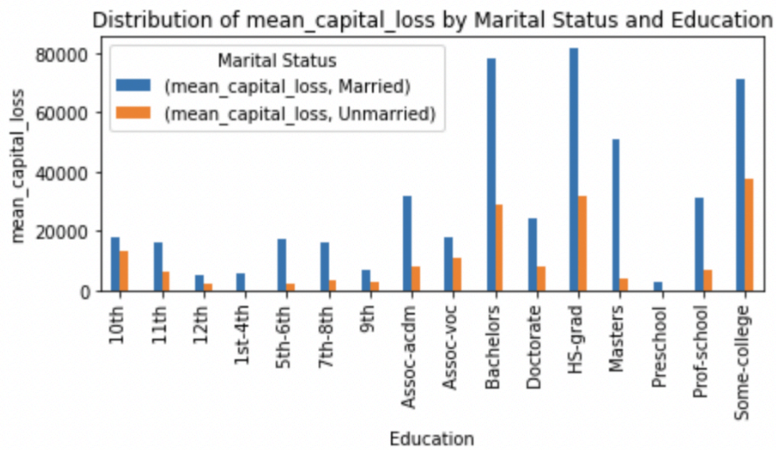
Occupation	Married	Unmarried
Adm-clerical	45972.99409684052	14425.247544961672
Armed-Forces	1887.0	0.0
Craft-repair	49862.656934907296	15045.589161503767
Exec-managerial	76006.94633410177	16296.548560587962
Farming-fishing	14273.317957981662	2283.688888888889
Handlers-cleaners	15181.70370631011	13506.044858523119
Machine-op-inspct	25914.987789877392	9598.036743164408
Other-service	30687.330729954032	8445.25309373538
Priv-house-serv	1594.0	1602.0
Prof-specialty	91075.86263238935	40651.08076929953
Protective-serv	12291.762155388471	7594.333333333333
Sales	50509.3566623556	15702.944196576404
Tech-support	27418.703402445506	9987.32302631579
Transport-moving	32170.003091998828	10606.346889952154

● This visualization displays the mean of capital loss for married and unmarried individuals, further divided by occupation.
● It effectively demonstrates the contrast in capital losses between married and unmarried individuals highlighting that mean of capital_loss is 0 for unmarried armed forces and highest in married Prof-speciality.



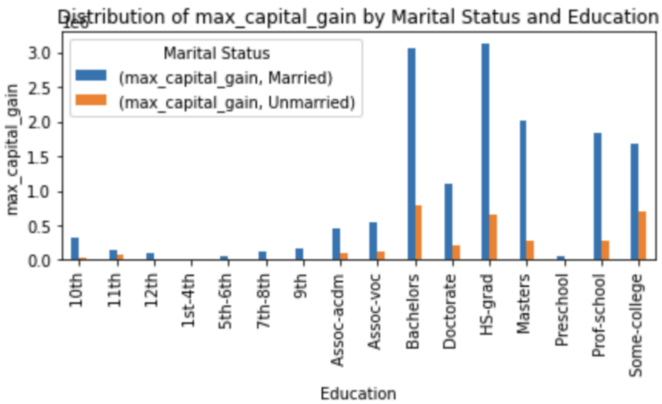
Education	Married	Unmarried
10th	31184	17164
11th	30518	17235
12th	7336	6666
1st-4th	8119	0
5th-6th	20389	2339
7th-8th	35362	5409
9th	9121	5784
Assoc-acdm	61424	17501
Assoc-voc	58073	18332
Bachelors	195456	74457
Doctorate	50836	12888
HS-grad	271670	109636
Masters	119835	20478
Preschool	3391	0
Prof-school	51810	12440
Some-college	217268	90938

- This visualization illustrates the maximum capital losses among both married and unmarried individuals, segmented by education level.
- It effectively emphasizes the differences in capital losses between the two marital status groups, offering valuable insights into their respective financial dynamics.
- Notably, unmarried individuals show zero capital losses in the 'Pre-school' and '1st-4th standard' education categories, while married individuals, particularly those with a high school or bachelor's degree, exhibit the highest capital losses.
- Overall, married individuals consistently demonstrate higher capital losses across all education categories.
- Also, we can observe that as the education category increases(example : the max capital losses in ‘1-4rth’ and ‘5th-6th’ have maximum capital losses increasing) , the maximum capital losses increases too until a certain level of education.



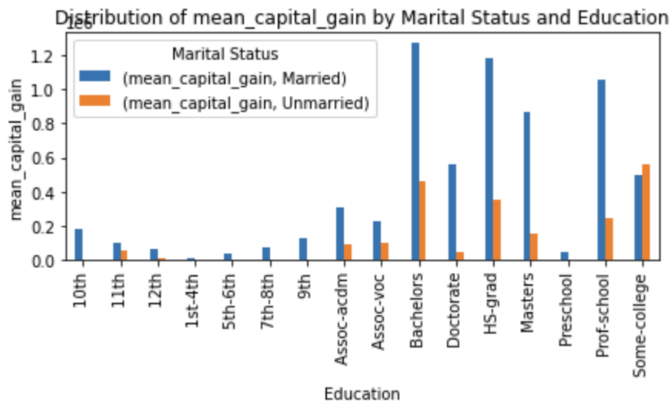
Education	Married	Unmarried
10th	17629.596623376623	12978.906666666666
11th	16392.002442528734	6445.9785526068135
12th	5381.5	2366.1815476190477
1st-4th	5516.25	0.0
5th-6th	17264.25	2339.0
7th-8th	16160.331746031747	3406.5
9th	6637.4	2662.2
Assoc-acdm	31864.8267588834206	8109.191666666667
Assoc-voc	17880.34421279289	10650.227272727272
Bachelors	78406.49170003516	28648.608111287598
Doctorate	24415.01884057971	8252.595238095239
HS-grad	81724.54177175486	31797.906266343824
Masters	50809.31627420761	3683.208785036102
Preschool	2555.0	0.0
Prof-school	31040.475300521077	7039.225
Some-college	71169.27982388793	37364.70795979317

- The visual presentation illustrates the average capital losses for married and unmarried individuals, categorized by their level of education.
- It presents a comparative analysis of mean capital losses across different educational levels and marital statuses, revealing intriguing trends. For instance, married individuals with a high school diploma or bachelor's degree tend to experience the highest capital losses, whereas unmarried individuals with some college education exhibit the highest losses.
- This visualization provides valuable insights into the relationship between education and marital status, highlighting potential differences in capital losses based on one's marital status and educational level.



Education	Married	Unmarried
10th	316901	42303
11th	150462	74766
12th	90088	16238
1st-4th	17473	3674
5th-6th	56439	2176
7th-8th	117036	6713
9th	162198	9632
Assoc-acdm	463847	102822
Assoc-voc	547178	121041
Bachelors	3067337	790558
Doctorate	1099688	213275
HS-grad	3136690	652792
Masters	2021222	273720
Preschool	45818	0
Prof-school	1830426	287340
Some-college	1672525	694500

- This visualization depicts the maximum capital gain recorded for married and unmarried individuals, further divided by education.
- The capital gain is higher among higher education levels like Bachelors, HD graduates , prof school and some college graduates in booth married and unmarried.



Education	Married	Unmarried
10th	177308.12307731694	5391.311872909699
11th	101156.09107889236	58312.68512634708
12th	66940.83333333333	12167.8
1st-4th	14448.2	1837.0
5th-6th	40488.4	1088.0
7th-8th	70967.19589999373	4280.5
9th	128901.38754578755	5307.833333333333
Assoc-acdm	304275.3287330903	94321.98333333334
Assoc-voc	230680.19390143332	99029.99444444444
Bachelors	1269323.1732117655	460571.4595389157
Doctorate	556970.8662902315	47374.2380952381
HS-grad	1178875.3437751618	352051.49108725885
Masters	865068.0642550108	156139.94568585887
Preschool	45818.0	0.0
Prof-school	1058106.4403776973	240397.6
Some-college	492718.9903259526	557835.8317027965

- The visual representation displays the mean capital gain by married and unmarried individuals, categorized by education.
- The mean capital gain is always higher in married among most of the education categories while in ‘some college’ category the mean capital gain is more in unmarried tha

Additional Work: (Extra credit)

“Combining multiple dimension attributes in one GROUP BY”

Please find the below code snippet in our implementation:

```python
def efficient_aggregate_sql(data, attributes_numeric, attributes_categorical,key):
    # Create a connection to an in-memory SQLite database
    conn = sqlite3.connect('census_data.db')
    # Create a cursor object
    cursor = conn.cursor()

    # Construct the SQL query for efficient aggregation
    numeric_aggregates = ', '.join([f'AVG({attributes_numeric[i]}) AS mean_{attributes_numeric[i]}, SUM({attributes_
    categorical_attributes = ', '.join(attributes_categorical)
    query =  f"""
        SELECT
            {categorical_attributes},
            {numeric_aggregates}
        FROM census_df
        GROUP BY """ + ','.join(attributes_categorical)+";"


    # Execute the SQL query
    cursor.execute(query)
    # Fetch the results
    results = cursor.fetchall()
    # Convert results to DataFrame
    columns = attributes_categorical + [f'{agg}_{attributes_numeric[i]}' for i in range(len(attributes_numeric)) for
    aggregated_data = pd.DataFrame(results, columns=columns)
```

Conclusion:

In conclusion, we successfully reproduced the approach and evaluation methodology outlined in the paper, achieving similar results in terms of identifying optimized aggregate views for married and unmarried individuals using the census dataset. The process involved data preprocessing, algorithm implementation, and result interpretation, contributing to the reproducibility of the original study.

Future Work:

Future work could focus on further optimizing the algorithms for larger datasets and exploring alternative utility measures for evaluating aggregate views. Additionally, conducting sensitivity analysis to assess the robustness of the results to parameter variations would enhance the reliability of the findings.