MIS272 – Predictive Analytics

T2 2024

Assignment 2 – Individual
Student name: Thi Ngoc Hong Nguyen
Student number: 222052067

Executive summary (1 page)

The company aims to optimize compensation structures by analysing data on employee salaries and benefits. A dataset with nearly 120K data points is given to identify high-paying job roles, predict total compensation based on key attributes, and explore relationships between job families and unions. These insights will guide data-driven decisions and improve financial management, employee satisfaction, and overall operational efficiency.

Objectives/Anticipated Benefits:

- Increase accuracy in predicting employee compensation, improving budgeting and planning.
- Identify high-value job roles to ensure competitive and fair compensation adjustments.
- Improve union negotiations by better understanding the relationships between job families and union associations.

Recommendations:

Based on created visualisations as to explore aspects of dataset:

- Company can prioritize Transit Operator and Special Nurse roles which have the highest average salaries and focus on maintaining competitive compensation for these roles to retain talent while addressing the low salary of Public Service Trainees (Figure 5)
- Company can balance non-salary benefits by enhancing non-salary benefits for General City Responsibilities and review compensation structures in Human Welfare & Neighborhood Development to maintain high benefits while adjusting sectors with lower non-salary components (Figure 8)
- Company can create strategies to address disparities in high-paying roles with lower-than-expected benefits and enhance retirement packages for lower-paid jobs to promote equity: Alignment of retirement benefits with salary levels (Figure 9)
- Company can enhance benefits for temporary employees to improve satisfaction and retention: Permanent employees receive the highest health and dental benefits, while temporary staff receive notably fewer (Figure 10)

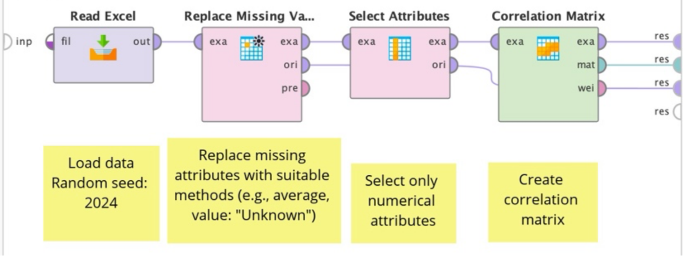
Based on predictive models:

- Linear Regression model: company should employ and regularly update the most productive model which has number of folds equal 10 with lower root mean squared error (6901.260) and high squared correlation (0.992) for salary planning
- Cluster model: company should implement the model with k=2 due to having the lowest Davies Bouldin Index (0.561) and shows that this configuration provided the most well-separated clusters
- FP-Growth model: company should leverage strong job-union relationships between Journeyman Trade and Public Safety Inspection; Journeyman Trade, Skilled Labor, and Supervisory-Labor & Trade with high lift and confidence. These relationships should guide union negotiations and help standardize compensation packages.

Data exploration, pattern discovery, and preparation (2 pages)

Preparation:

- Load Data (Random Seed: 2024): The dataset contains nearly 120,000 data points requiring careful handling to ensure consistency and accuracy throughout analysis
- Replace Missing Values: Missing values for Job, Union, and Employment Type are replaced with "Unknown"; for Union Code (numerical attributes), missing values are replaced with the average values to ensure the central trend of the data is maintained without introducing bias and maintain the dataset's comprehensiveness.
- Select Attributes: Select only numerical attributes (e.g., Salary, Total Benefits, Health and Dental,..) for creating a correlation matrix. It is used to identify any strong relationships between variables so as to guide further exploration in subsequent analyses



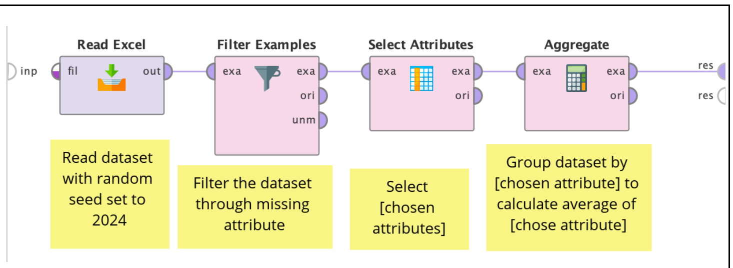
Load data Random seed: 2024

Replace missing attributes with suitable methods (e.g., average, value: "Unknown")

Select only numerical attributes

Create correlation matrix

Figure 1: Preparation



Read dataset with random seed set to 2024

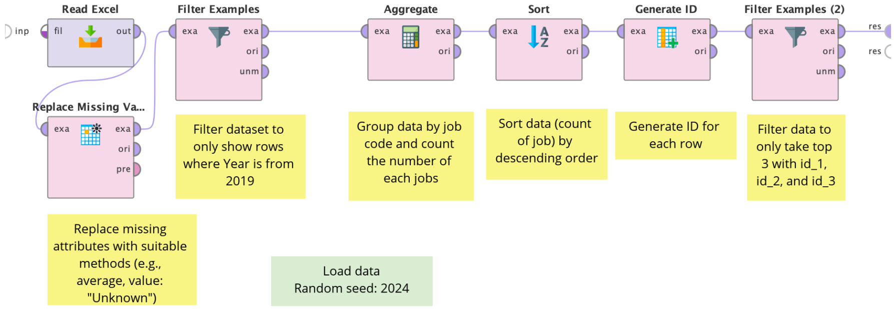
Filter the dataset through missing attribute

Select [chosen attributes]

Group dataset by [chosen attribute] to calculate average of [chose attribute]

Figure 2: Process for visualisation 3 & 4

Visualisation



Filter dataset to only show rows where Year is from 2019

Group data by job code and count the number of each jobs

Sort data (count of job) by descending order

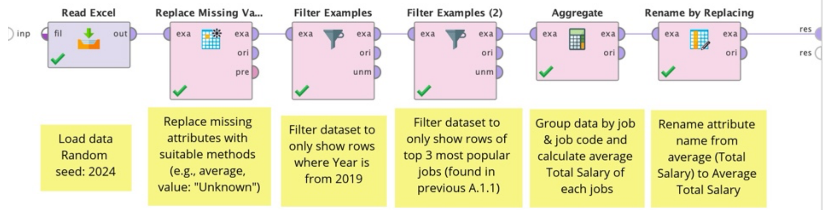
Generate ID for each row

Filter data to only take top 3 with id_1, id_2, and id_3

Replace missing attributes with suitable methods (e.g., average, value: "Unknown")

Load data Random seed: 2024

Figure 3: Process for finding top 3 most common jobs



Load data Random seed: 2024

Replace missing attributes with suitable methods (e.g., average, value: "Unknown")

Filter dataset to only show rows where Year is from 2019

Filter dataset to only show rows of top 3 most popular jobs (found in previous A.1.1)

Group data by job & job code and calculate average Total Salary of each jobs

Rename attribute name from average (Total Salary) to Average Total Salary

Figure 4: Process for calculating average total salary of top 3 most common jobs

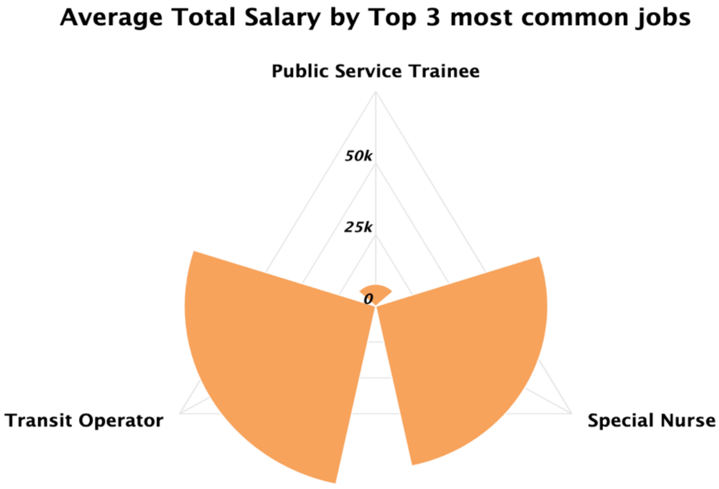


Figure 5: Average Total Salary by top 3 most common jobs

Interpretation

Transit Operator has the highest average total salary ($63488), followed by Special Nurse ($57010), and Public Service Trainee only earns $7695

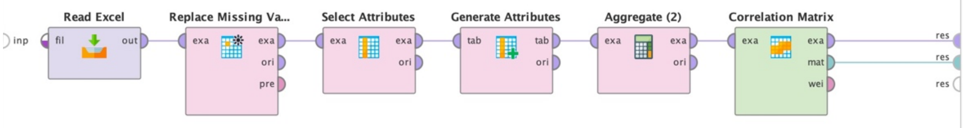
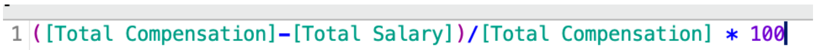


Figure 6: Process



```
([Total Compensation]-[Total Salary])/[Total Compensation] * 100
```

Figure 7: Function expression to calculate Disparity Percentage

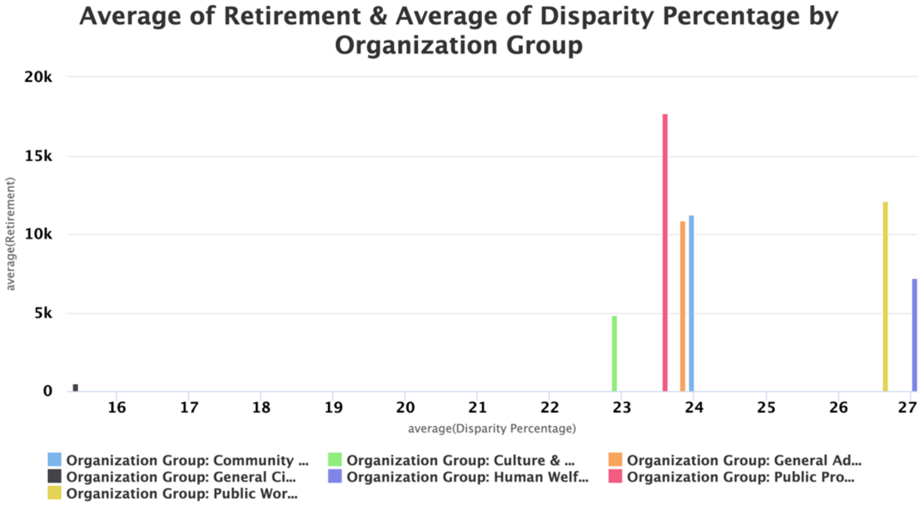


Figure 8: Average Retirement & Average Disparity Percentage by Organization Group

- Human Welfare & Neighborhood Development has the highest disparity (26.96%) which shows a larger portion of compensation comes from non-salary benefits
- General City Responsibilities has the lowest disparity (15.43%) and focus more on salary with minimal retirement benefits
- Public Protection offers the highest retirement benefits ($17,675.84) with a moderate disparity percentage (23.42%)

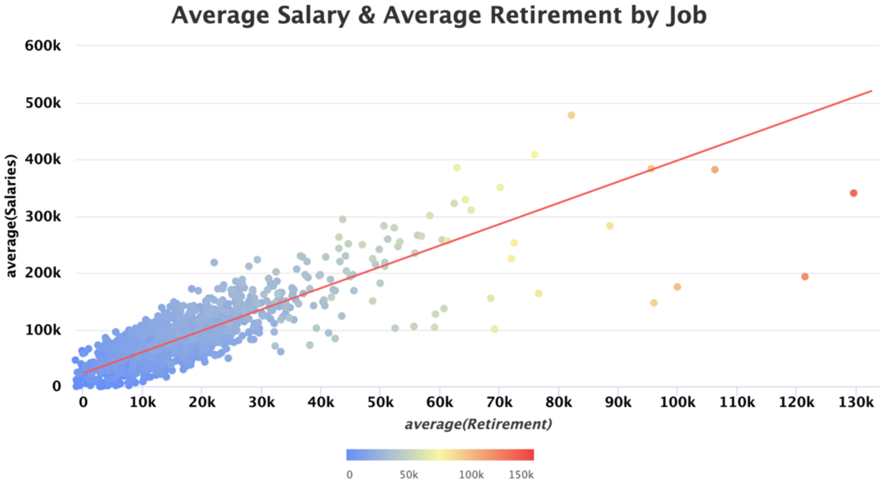


Figure 9: Average Salary & Average Retirement by Job

The scatter points show some variation with a few high-paying jobs having slightly lower retirement averages compared to their salaries. However, the general trend indicates that higher-paid jobs often correlate with higher retirement benefits

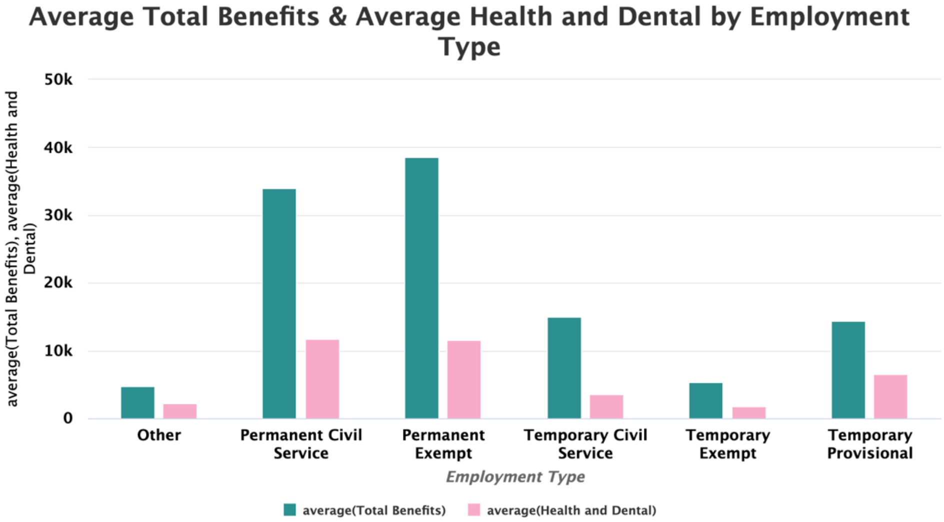


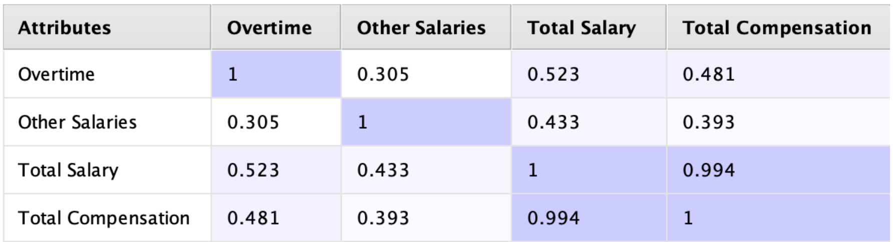
Figure 10: Average Total Benefits & Average Health and Dental by Employment Type

- Permanent employees, particularly those classified as Permanent Exempt receive the highest total benefits which is significantly more than any other category.
- In contrast, temporary employees, especially those in the Temporary Exempt category receive notably fewer health and dental benefits.
- The "Other" category has the lowest benefits across both metrics.

Predictive modelling (2 pages)

1. Linear Regression – Cross Validation

- *Data Preparation:*
  - Select attribute: Select Other Salaries, Overtime, Total Compensation, Total Salary to be used in this model. These attributes are carefully chosen after remove all multicollinearity between other independent variables, thereby avoiding any distortion in the relationship with the target variable.



Attributes	Overtime	Other Salaries	Total Salary	Total Compensation
Overtime	1	0.305	0.523	0.481
Other Salaries	0.305	1	0.433	0.393
Total Salary	0.523	0.433	1	0.994
Total Compensation	0.481	0.393	0.994	1

Figure 11: Final Correlation Matrix

- Set Role: Use Total Compensation as the label for the prediction model

By using 10-fold, each model is tested on smaller 10% subsets of the data which might reduce bias and variability between folds to make the performance metric more stable and reliable. However, this dataset has large sizing (120K data), 3 folds is also used to value computational efficiency and speed. Besides, to evaluate the accuracy of the model's predictions, "Generate Attribute" operator calculates the residual for "Total Compensation" by subtracting the predicted value from the actual value to show the difference or error in predictions. In terms of Performance, this model might calculate root mean squared error and squared correlation to evaluate the model's Total Compensation prediction ability.

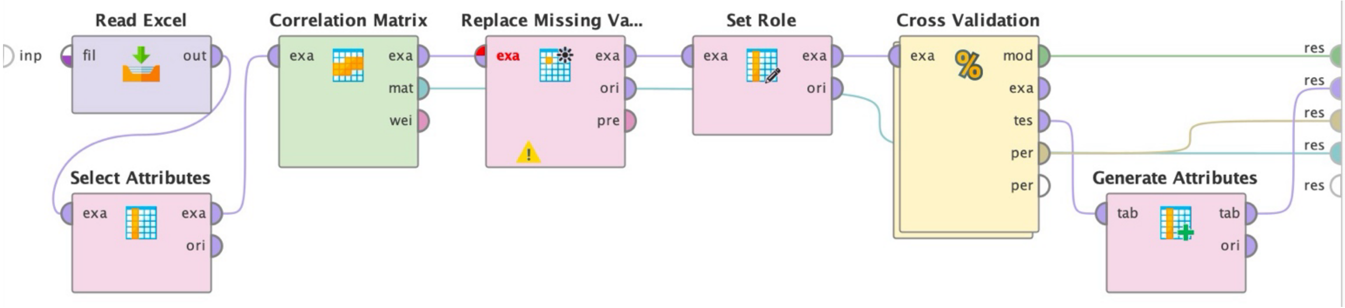


Figure 12: Linear Regression - Cross Validation process

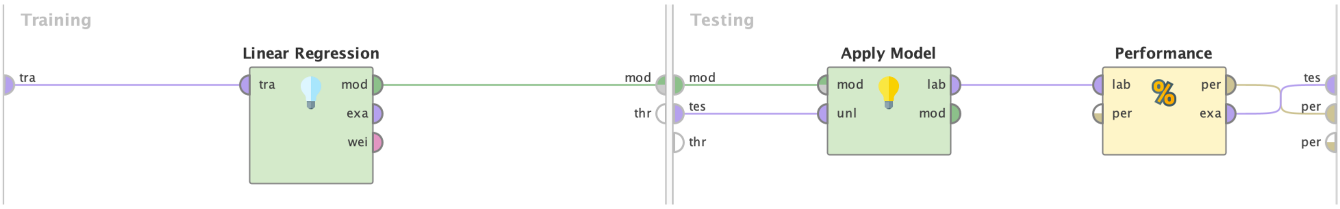


Figure 13: Operators inside Cross Validation

1. Cluster Model (Extension)

- *Data Preparation:*
  - Select Attributes: Focus on numerical attributes such as Health and Dental, Other Benefits, Retirement, Total Benefits, Total Salary, and Total Compensation to ensure relevant clustering
  - Normalize: Scale the numerical attributes within a set range (e.g., 0 to 1) to prevent discrepancies caused by differing attribute scales

The clustering operator use:

- K: Set to 2, 3, and 4 clusters to balance simplicity and detail because too few clusters might miss complexity while too many may overfit
- Max runs: 10 iterations limit might ensure efficiency with large datasets by stopping unnecessary processing
- Bregman Divergences: Provides adaptability to different data types
- Squared Euclidean Distance: Focuses on separating data points and helps identify distinct clusters
- Max optimization steps: 100 steps might allow enough refinement without over-optimizing and maintain efficiency

The Performance (Cluster Distance Performance) operator evaluates clustering by measuring the average distance from each point to its cluster centroid. Besides, the maximize setting ensures the algorithm seeks to optimize cluster tightness.

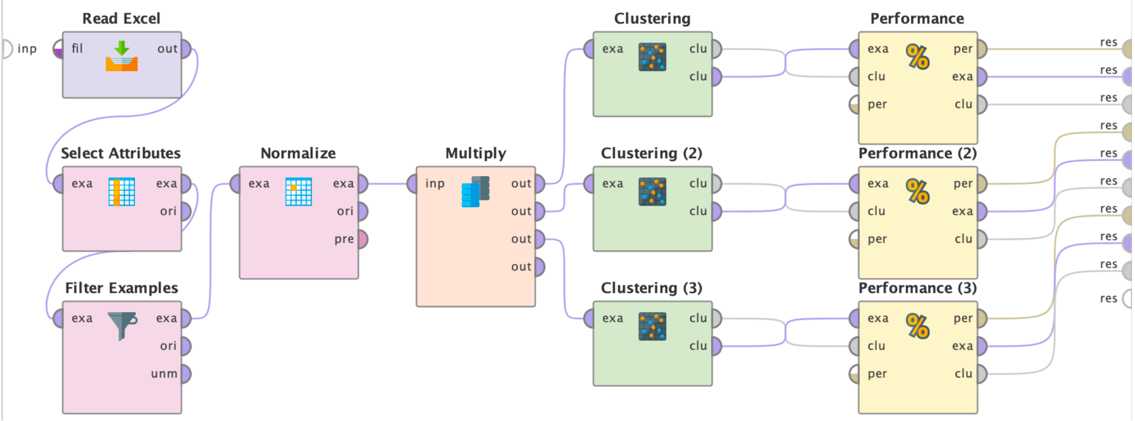


Figure 14: Cluster Model

3. Association Rule mining using the FP-Growth

- *Data Preparation:*
  - Pivot: Group data by Union and sums the Counting column to capture and summarize job-family relationships for further rule mining
  - Replace Missing Values: Fill missing data with zero values to ensure completeness and ready the dataset for analysis
  - Numerical to Binominal: Convert numerical fields into binary values to make the data suitable for the FP-Growth algorithm
  - Select attribute: sum(Counting)_
  - Rename by Replacing: rename all attributes to make them more intuitive and relevant for analysis which ensures consistency and clarity in the dataset
- *In the FP-Growth operator:*
  - Min support (0.3): Ensure that only item sets appearing in at least 30% of transactions are considered, filtering out insignificant sets and focus only on those that have strong support
  - Min items per itemset (1): Guarantees each itemset has at least one item to avoid empty sets
  - Max items per itemset (0): No upper limit allows for large combinations of items
  - Max number of itemsets (1,000,000): Caps the number of generated itemsets to manage memory and efficiency
  - Min number of itemsets (30): Ensures that the model does not stop early by forcing it to find at least 30 frequent itemsets
  - Max retries (15): Provides flexibility for the model to find frequent sets
  - Requirement decrease factor (0.9): Allows for progressively lowering support thresholds to find more itemsets if initial levels are too high

Regarding Create Association Rules, this model will select lift as the main metric for rule generation, choose 0.8 as the min criterion value so lift value lower than 0.8 will not be considered; thus, to find interesting patterns while avoiding weak or insignificant associations. Moreover, because this data likely contains many common job-union relationships, taking gain theta at 2.0 ensures focusing on those that show considerable improvement in associations. Finally, laplace k set to 1.0 helps avoid overfitting in cases with sparse data or small support.

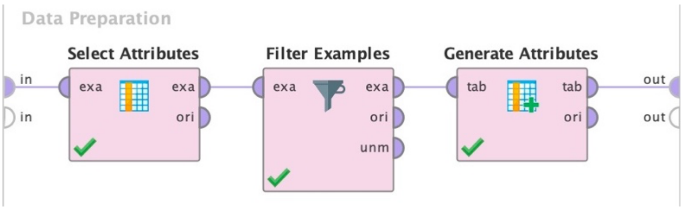


Figure 15: Operators inside Subprocess (Data Preparation)

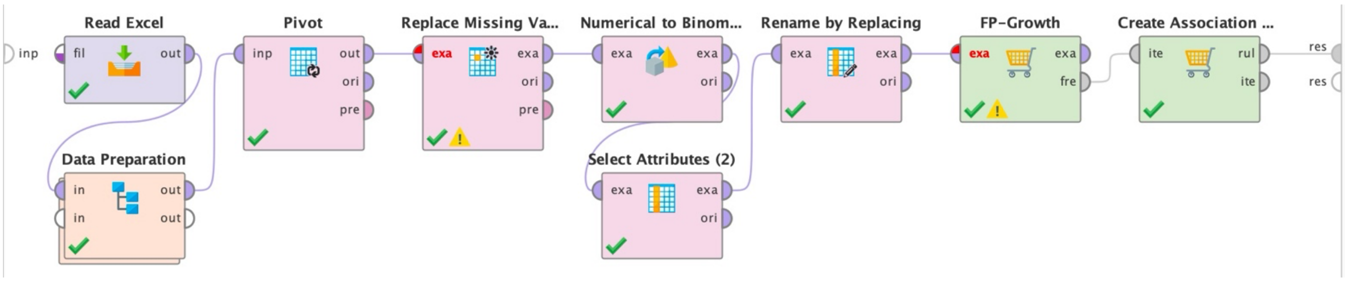


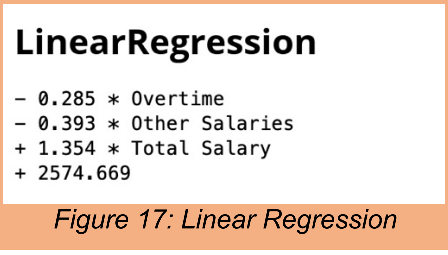
Figure 16: Process association rule mining using the FP-Growth

Model evaluation and improvement (2 pages)

RESULT TABLE

Predictive Modelling

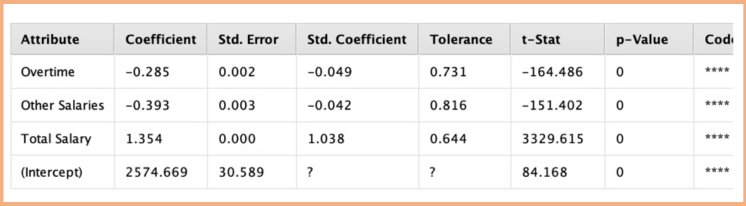
Linear Regression – Cross Validation



```
LinearRegression

− 0.285 ∗ Overtime
− 0.393 ∗ Other Salaries
+ 1.354 ∗ Total Salary
+ 2574.669
```

Figure 17: Linear Regression



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Cod
Overtime	-0.285	0.002	-0.049	0.731	-164.486	0	****
Other Salaries	-0.393	0.003	-0.042	0.816	-151.402	0	****
Total Salary	1.354	0.000	1.038	0.644	3329.615	0	****
(Intercept)	2574.669	30.589	?	?	84.168	0	****

Figure 18: Linear Regression

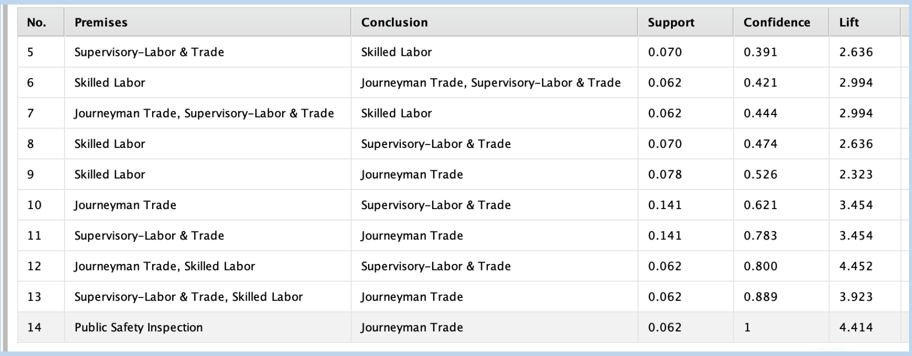
Parameter	Root mean squared error	Squared correlation
Number of folds = 3	6902.108 ± 70.228	0.992
Number of folds = 10	*6901.260 ± 113.429*	*0.992*

Cluster Model (Extension)

Parameter	Average within-centroid distance	Davies Bouldin
K = 2	*0.017*	*0.561*
K = 3	0.012	0.855
K = 4	0.010	0.988

Association Rule mining using the FP-Growth

1. ***Min support: 0.3***



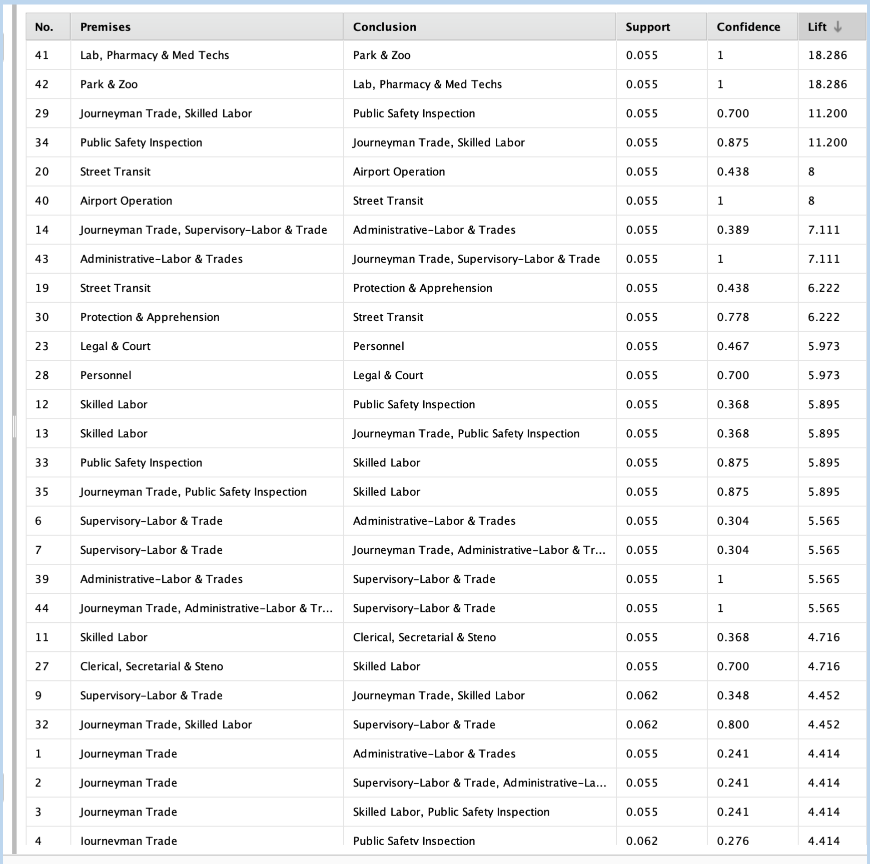
No.	Premises	Conclusion	Support	Confidence	Lift
5	Supervisory-Labor & Trade	Skilled Labor	0.070	0.391	2.636
6	Skilled Labor	Journeyman Trade, Supervisory-Labor & Trade	0.062	0.421	2.994
7	Journeyman Trade, Supervisory-Labor & Trade	Skilled Labor	0.062	0.444	2.994
8	Skilled Labor	Supervisory-Labor & Trade	0.070	0.474	2.636
9	Skilled Labor	Journeyman Trade	0.078	0.526	2.323
10	Journeyman Trade	Supervisory-Labor & Trade	0.141	0.621	3.454
11	Supervisory-Labor & Trade	Journeyman Trade	0.141	0.783	3.454
12	Journeyman Trade, Skilled Labor	Supervisory-Labor & Trade	0.062	0.800	4.452
13	Supervisory-Labor & Trade, Skilled Labor	Journeyman Trade	0.062	0.889	3.923
14	Public Safety Inspection	Journeyman Trade	0.062	1	4.414

Figure 19: Result of FP-Growth

- Both confidence and lift are in top the highest:
  - Journeyman Trade, Skilled Labor frequently associate with Supervisory-Labor & Trade (at about 0.8 and 4.452 respectively)
  - Supervisory-Labor & Trade, Skilled Labor is frequently listed with Journeyman Trade (0.889 and 3.923 respectively)
  - Journeyman Trade and Public Safety Inspection are strongly associated (Confidence: 1.0 – Lift: 4.4)
- Journeyman Trade and Supervisory-Labor & Trade frequently associate with Skilled Labor (Confidence: 0.526 & 0.474)
- Supervisory-Labor & Trade is frequently listed with Journeyman Trade (Confidence: 0.783, 0.621 – Lift: 3.454)

- Rules Found: Fewer rules generated compared to lower support
- Confidence: The confidence values range up to 1.0 but many rules show confidence below 0.5
- Lift: Lift values go up to 4.45 with stronger associations for some rules

2. Min support: 0.05



No.	Premises	Conclusion	Support	Confidence	Lift ↓
41	Lab, Pharmacy & Med Techs	Park & Zoo	0.055	1	18.286
42	Park & Zoo	Lab, Pharmacy & Med Techs	0.055	1	18.286
29	Journeyman Trade, Skilled Labor	Public Safety Inspection	0.055	0.700	11.200
34	Public Safety Inspection	Journeyman Trade, Skilled Labor	0.055	0.875	11.200
20	Street Transit	Airport Operation	0.055	0.438	8
40	Airport Operation	Street Transit	0.055	1	8
14	Journeyman Trade, Supervisory-Labor & Trade	Administrative-Labor & Trades	0.055	0.389	7.111
43	Administrative-Labor & Trades	Journeyman Trade, Supervisory-Labor & Trade	0.055	1	7.111
19	Street Transit	Protection & Apprehension	0.055	0.438	6.222
30	Protection & Apprehension	Street Transit	0.055	0.778	6.222
23	Legal & Court	Personnel	0.055	0.467	5.973
28	Personnel	Legal & Court	0.055	0.700	5.973
12	Skilled Labor	Public Safety Inspection	0.055	0.368	5.895
13	Skilled Labor	Journeyman Trade, Public Safety Inspection	0.055	0.368	5.895
33	Public Safety Inspection	Skilled Labor	0.055	0.875	5.895
35	Journeyman Trade, Public Safety Inspection	Skilled Labor	0.055	0.875	5.895
6	Supervisory-Labor & Trade	Administrative-Labor & Trades	0.055	0.304	5.565
7	Supervisory-Labor & Trade	Journeyman Trade, Administrative-Labor & Tr...	0.055	0.304	5.565
39	Administrative-Labor & Trades	Supervisory-Labor & Trade	0.055	1	5.565
44	Journeyman Trade, Administrative-Labor & Tr...	Supervisory-Labor & Trade	0.055	1	5.565
11	Skilled Labor	Clerical, Secretarial & Steno	0.055	0.368	4.716
27	Clerical, Secretarial & Steno	Skilled Labor	0.055	0.700	4.716
9	Supervisory-Labor & Trade	Journeyman Trade, Skilled Labor	0.062	0.348	4.452
32	Journeyman Trade, Skilled Labor	Supervisory-Labor & Trade	0.062	0.800	4.452
1	Journeyman Trade	Administrative-Labor & Trades	0.055	0.241	4.414
2	Journeyman Trade	Supervisory-Labor & Trade, Administrative-La...	0.055	0.241	4.414
3	Journeyman Trade	Skilled Labor, Public Safety Inspection	0.055	0.241	4.414
4	Journeyman Trade	Public Safety Inspection	0.062	0.276	4.414

Figure 20: Result of FP-Growth

- Rules Found: A greater number of rules are generated with lower support
- Confidence: There is a wider range of confidence values with many rules showing confidence above 0.7 and even reaching 1.0
- Lift: Lift values are higher, with some rules reaching up to 18.286, indicating very strong associations

Performance Improvements

- *Linear Regression Model:* Increasing the number of folds in cross-validation from 3 to 10 slightly reduced the RMSE (from 6902.108 ± 70.228 to 6901.260 ± 113.429) shows a marginal improvement in prediction accuracy. Besides, the squared correlation remained high at 0.992 in both cases which indicates that the model maintains strong predictive power across both configurations.
- *Clustering Model:* Employing various quantities of clusters (k = 2, k = 3, and k = 4) in the clustering analysis showed varying levels of cluster distinctiveness. Noticeably, the Davies Bouldin Index was lowest at k = 2 (0.561) and shows that this configuration provided the most well-separated clusters. As the number of clusters increased to 4, the Davies Bouldin Index rose to 0.988 which suggests that the clusters became less distinct.
- *The FP-Growth model:* Demonstrated noticeable improvements by adjusting the minimum support. Specifically, lowering the min support from 0.3 to 0.05 increased the rule set, captured rarer associations and raised the maximum lift to 18.286.

Conclusion

- *Linear Regression Model:* While increasing the number of folds in cross-validation led to a small improvement in RMSE, the model's performance was generally stable with minimal variance between different fold settings. Thus, the model provides reliable predictions for salary components and make it well-suited for supporting decision-making in employee compensation strategies.
- *Clustering Model:* The clustering model with k = 2 yielded the best performance which offers the clearest and most distinct groupings. Besides, this configuration would be most beneficial for categorizing employees based on their compensation and benefits. However, increasing the number of clusters might provide additional granularity depending on the business need for more detailed segmentation.
- *The FP-Growth model:* When evaluating the models, it is important to focus on both support and confidence, but in this case, lift played a significant role in identifying the strongest associations. For instance, the lower minimum support value of 0.05 enabled the discovery of less common but stronger associations as indicated by the high lift. Therefore, although the min support of 0.05 provides more in-depth analysis, the min support of 0.3 directly answers to question of configuration for uncovering stronger relationships between Job Families and Unions.