Dataset description

October 22, 2024

1 Dataset Description Report

1.1 Group Members

- Ruslan Basyrov (12209898)
- Iana Bembeeva (52017248)
- Ekaterina Grigorashchenko (12432933)

1.2 Introduction

This report provides a detailed description of two datasets used for analysis: a weather dataset predicting rainfall and a salary dataset. Each dataset's characteristics, attributes, and significance are discussed.

1.3 Dataset 1: Weather Observations Dataset

1.3.1 Context

The first dataset comprises daily weather observations collected over approximately ten years from various locations across Australia. This dataset enables the prediction of next-day rain, answering the crucial question of whether to carry an umbrella. Link to Kaggle.

1.3.2 Characteristics

- **Samples**: Approximately 10 years of daily observations (exact count of samples may vary).
- **Attributes**: 22 attributes, including both numerical and categorical types.

1.3.3 Attribute Types and Unique Values

Attribute	Type	Unique Values	Missing Values	Description
Date	Ordinal	3436	0	The date of observation.
Location	Nominal	49	0	The name of the weather station location.
MinTemp	Ratio	389	1485	Minimum temperature in degrees Celsius.
MaxTemp	Ratio	505	1261	Maximum temperature in degrees Celsius.
Rainfall	Ratio	681	3261	Amount of rainfall recorded for the day in mm.
Evaporation	Ratio	358	62790	Class A pan evaporation (mm) in the 24 hours to 9am.

Attribute	Type	Unique Values	Missing Values	Description
Sunshine	Ratio	145	69835	Number of hours of bright sunshine in the day.
WindGustDir	Nominal	16	10326	Direction of the strongest wind gust in the last 24 hours.
WindGustSpeed	Ratio	67	10263	Speed (km/h) of the strongest wind gust in the last 24 hours.
WindDir9am	Nominal	16	10566	Direction of the wind at 9am.
WindDir3pm	Nominal	16	4228	Direction of the wind at 3pm.
WindSpeed9am	Ratio	43	1767	Wind speed (km/h) averaged over 10 minutes prior to 9am.
WindSpeed3pm	Ratio	44	3062	Wind speed (km/h) averaged over 10 minutes prior to 3pm.
Humidity9am	Ratio	101	2654	Humidity (percent) at 9am.
Humidity3pm	Ratio	101	4507	Humidity (percent) at 3pm.
Pressure9am	Ratio	546	15065	Atmospheric pressure (hPa) at 9am.
Pressure3pm	Ratio	549	15028	Atmospheric pressure (hPa) at 3pm.
Cloud9am	Ratio	10	55888	Fraction of sky obscured by cloud at 9am (in oktas).
Cloud3pm	Ratio	10	1767	Fraction of sky obscured by cloud at 3pm (in oktas).
Temp9am	Ratio	441	3609	Temperature (degrees C) at 9am.
Temp3pm	Ratio	502	1485	Temperature (degrees C) at 3pm.
RainToday	Nominal	2	3261	Yes if precipitation (mm) exceeds 1mm, otherwise No.
RainTomorrow	Nominal	2	3267	Indicates whether it will rain tomorrow (Yes/No).

1.3.4 Comments on Reformatting

- **Date**: Should be reformatted to `datetime` for better handling in time series analysis.

1.3.5 Target Attribute

- **RainTomorrow**: This binary attribute indicates whether it will rain the following day (Yes/No). It is crucial for predicting weather patterns and making informed decisions.

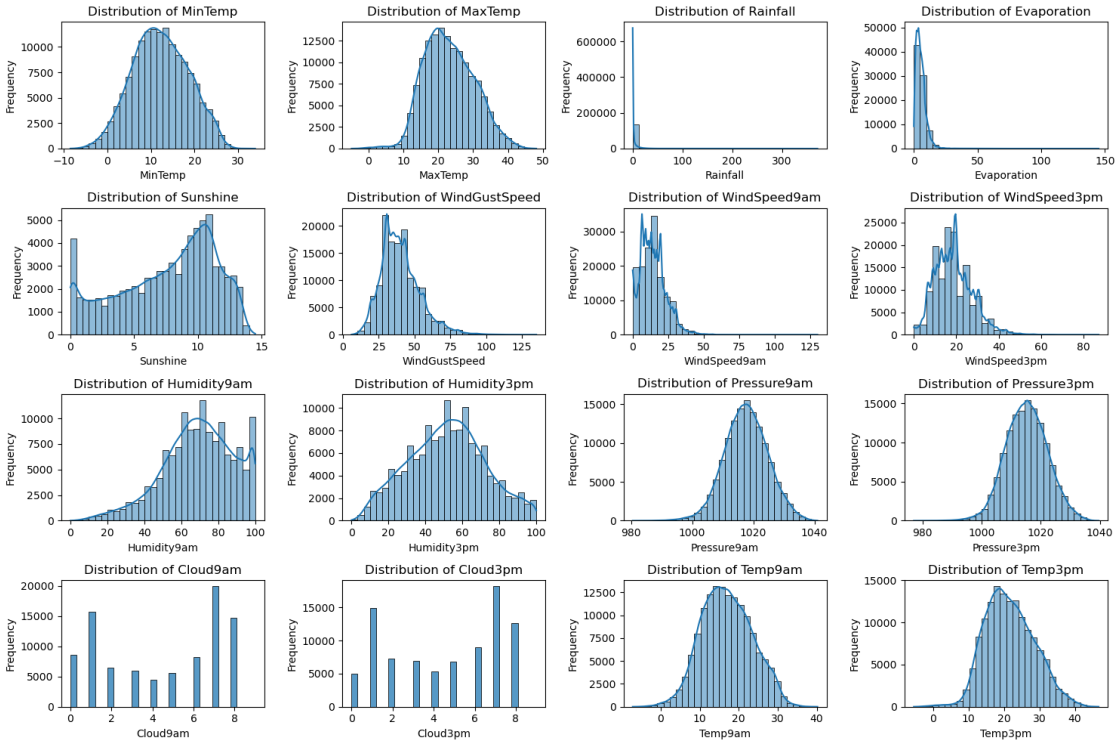
1.3.6 Importance of Dataset

Understanding the distribution of values in attributes helps in preprocessing steps, such as handling missing values and feature selection, to build accurate predictive models.

	0	1	2	3	4
Date	2008-12-01	2008-12-02	2008-12-03	2008-12-04	2008-12-05
Location	Albury	Albury	Albury	Albury	Albury
MinTemp	13.400000	7.400000	12.900000	9.200000	17.500000
MaxTemp	22.900000	25.100000	25.700000	28.000000	32.300000
Rainfall	0.600000	0.000000	0.000000	0.000000	1.000000
Evaporation	NaN	NaN	NaN	NaN	NaN
Sunshine	NaN	NaN	NaN	NaN	NaN
WindGustDir	W	WNW	WSW	NE	W
WindGustSpeed	44.000000	44.000000	46.000000	24.000000	41.000000
WindDir9am	W	NNW	W	SE	ENE
WindDir3pm	WNW	WSW	WSW	E	NW
WindSpeed9am	20.000000	4.000000	19.000000	11.000000	7.000000
WindSpeed3pm	24.000000	22.000000	26.000000	9.000000	20.000000
Humidity9am	71.000000	44.000000	38.000000	45.000000	82.000000
Humidity3pm	22.000000	25.000000	30.000000	16.000000	33.000000
Pressure9am	1007.700000	1010.600000	1007.600000	1017.600000	1010.800000
Pressure3pm	1007.100000	1007.800000	1008.700000	1012.800000	1006.000000
Cloud9am	8.000000	NaN	NaN	NaN	7.000000
Cloud3pm	NaN	NaN	2.000000	NaN	8.000000
Temp9am	16.900000	17.200000	21.000000	18.100000	17.800000
Temp3pm	21.800000	24.300000	23.200000	26.500000	29.700000
RainToday	No	No	No	No	No
RainTomorrow	No	No	No	No	No

1.3.7 Dataset visualizations

1. **Distribution of RainTomorrow** - understanding the balance between classes
2. **Histograms of Numeric Attributes**



1.4 Dataset 2: Employee Salary Dataset

1.4.1 Context

The second dataset provides annual salary information, including gross and overtime pay, for all active, permanent employees of Montgomery County, MD, for the calendar year 2016. This data is essential for analyzing salary distribution and identifying trends in employee compensation. Link to OpenML

1.4.2 Characteristics

- **Samples**: 9222 employee records
- **Attributes**: 13 attributes with a mix of numeric and categorical types.

1.4.3 Attribute Types and Unique Values

Attribute	Type	Unique Values	Description
full_name	Nominal	9222	Employee's full name.
gender	Nominal	2	Employee's gender (F, M).
current_annual_salary	Ratio	3403	Employee's current annual salary in USD.
2016_gross_pay_received	Ratio	8977	Total gross pay received in 2016.
2016_overtime_pay	Ratio	6176	Total overtime pay received in 2016.
department	Nominal	37	Department of the employee.
department_name	Nominal	37	Full name of the department.
division	Nominal	694	Division of the employee within the department.
assignment_category	Nominal	2	Employment category (Fulltime-Regular, Parttime-Regular).
employee_position_title	Nominal	385	Job title of the employee.
underfilled_job_title	Nominal	84	Job title of the position being underfilled, if applicable.
date_first_hired	Ordinal	2264	Date the employee was first hired.
year_first_hired	Interval	51	Year the employee was first hired.

1.4.4 Target Attribute

- **current_annual_salary**: The primary target attribute used for salary analysis and modeling. Understanding its distribution is crucial for various analyses, including equity and budgeting.

1.4.5 Comments on Reformatting

- **date_first_hired**: Should be reformatted to `datetime` for better handling in time analysis.

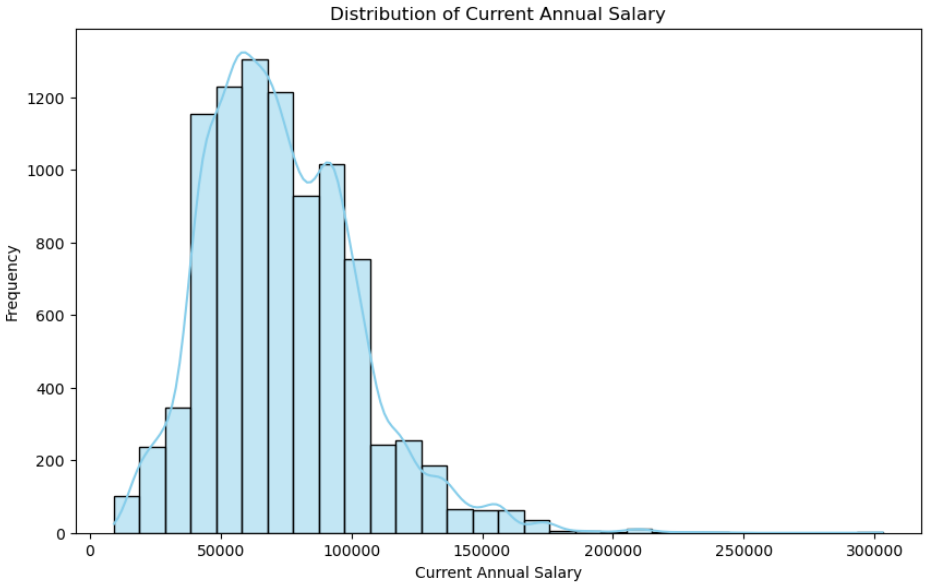
1.4.6 Importance of Dataset

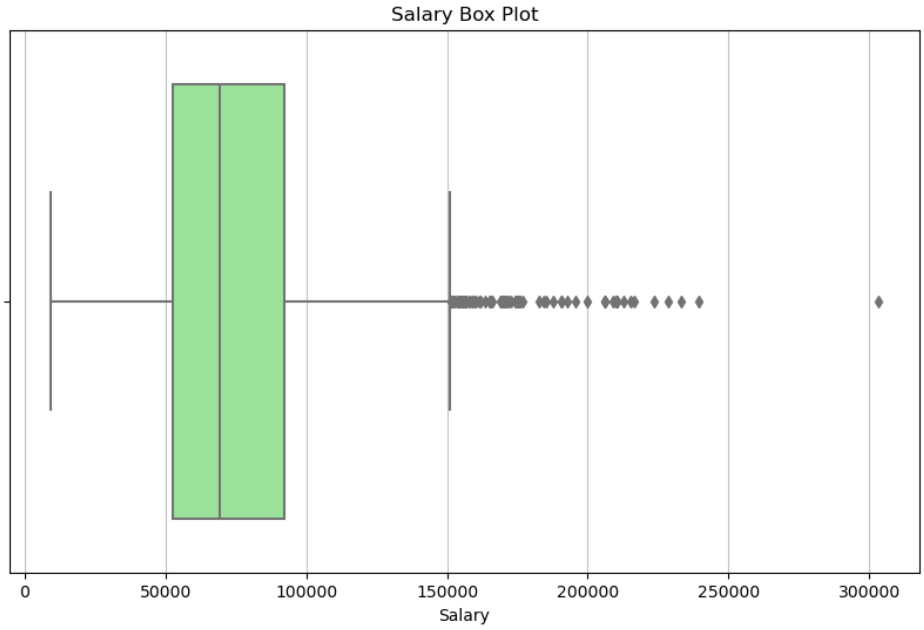
The distribution of numeric values in the salary dataset provides insight into compensation trends, while the categorical data (such as gender and department) allows for analyzing disparities and ensuring equitable pay practices.

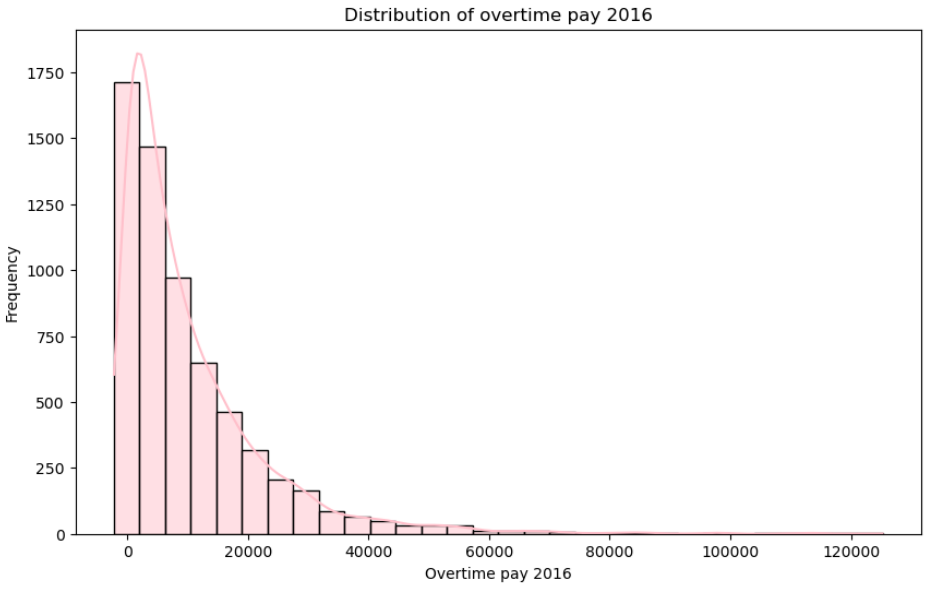
1.4.7 Dataset visualizations

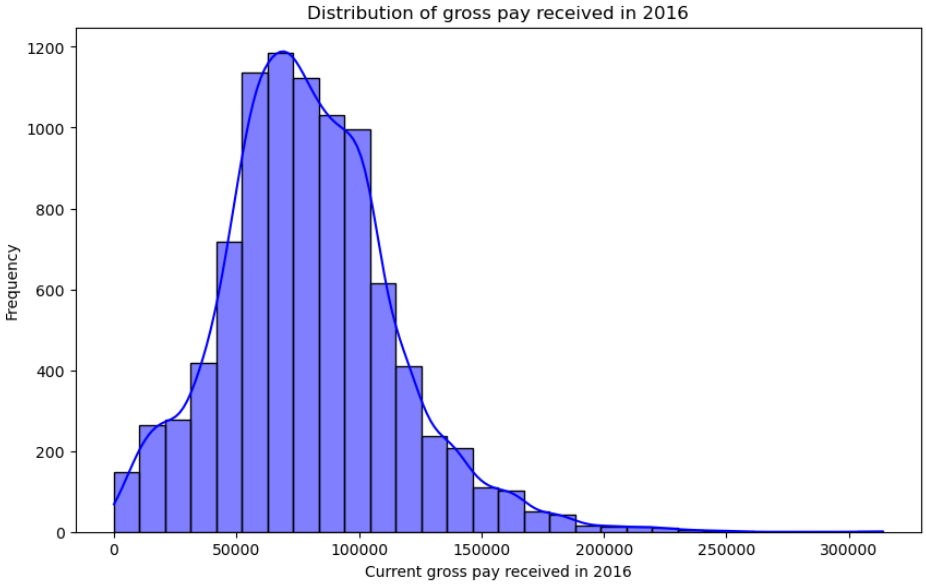
1. **Distribution of Current Annual Salary** - analyzing salary ranges and identifying any outliers

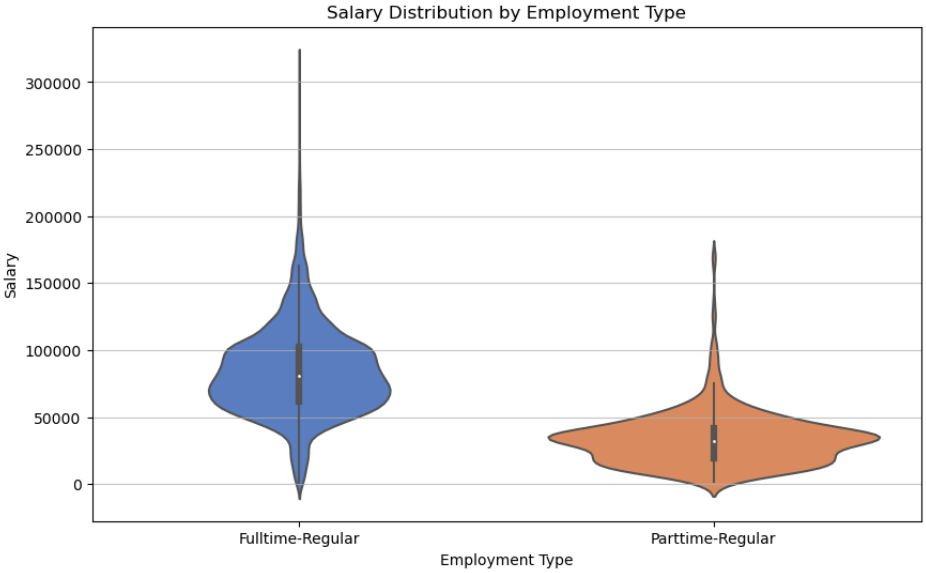
Attribute	Instance 1	Instance 2	Instance 3
full_name	Aarhus, Pam J.	Aaron, David J.	Aaron, Marsha M.
gender	F	M	F
current_annual_salary	69,222.18	97,392.47	104,717.28
2016_gross_pay_received	71,225.98	103,088.48	107,000.24
2016_overtime_pay	416.1	3,326.19	1,353.32
department	POL	POL	HHS
department_name	Department of Police	Department of Police	Department of Health...
division	MSB Information Mgmt...	ISB Major Crimes Divi...	Adult Protective and...
assignment_category	Fulltime-Regular	Fulltime-Regular	Fulltime-Regular
employee_position_title	Office Services Coord...	Master Police Officer	Social Worker IV
underfilled_job_title	None	None	None
date_first_hired	09/22/1986	09/12/1988	11/19/1989
year_first_hired	1986	1988	1989

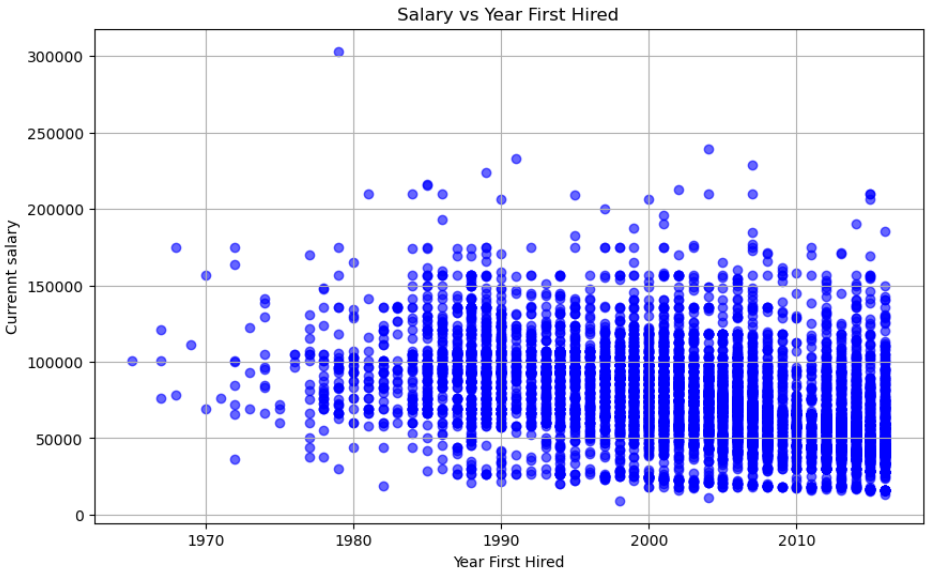



Salary Box Plot
0
50000
100000
150000
200000
250000
300000
Salary

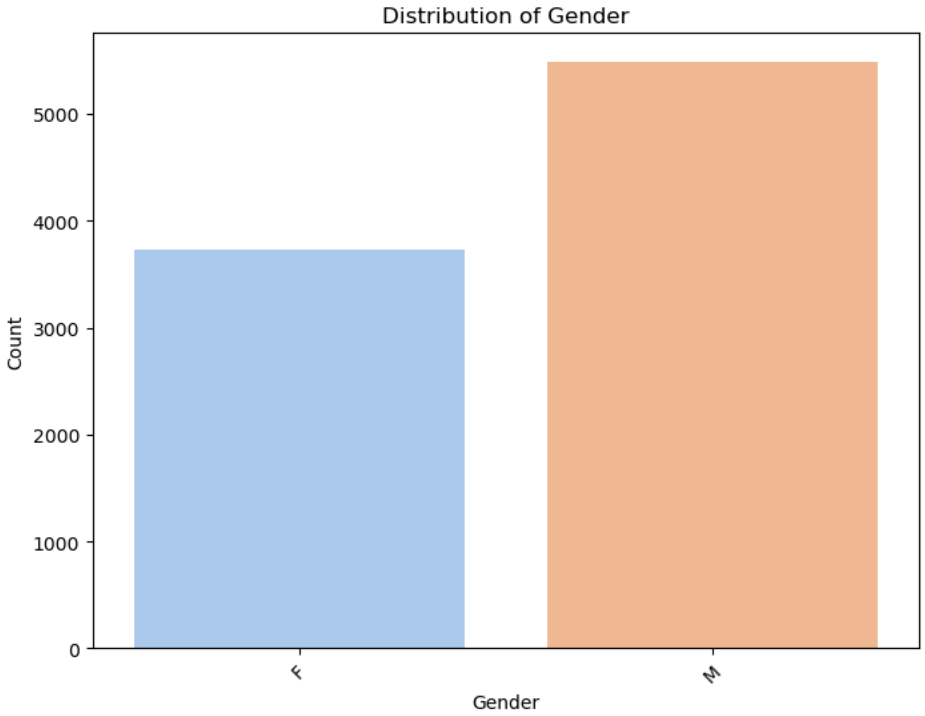

Distribution of overtime pay 2016
Frequency
0
250
500
750
1000
1250
1500
1750
0
20000
40000
60000
80000
100000
120000
Overtime pay 2016

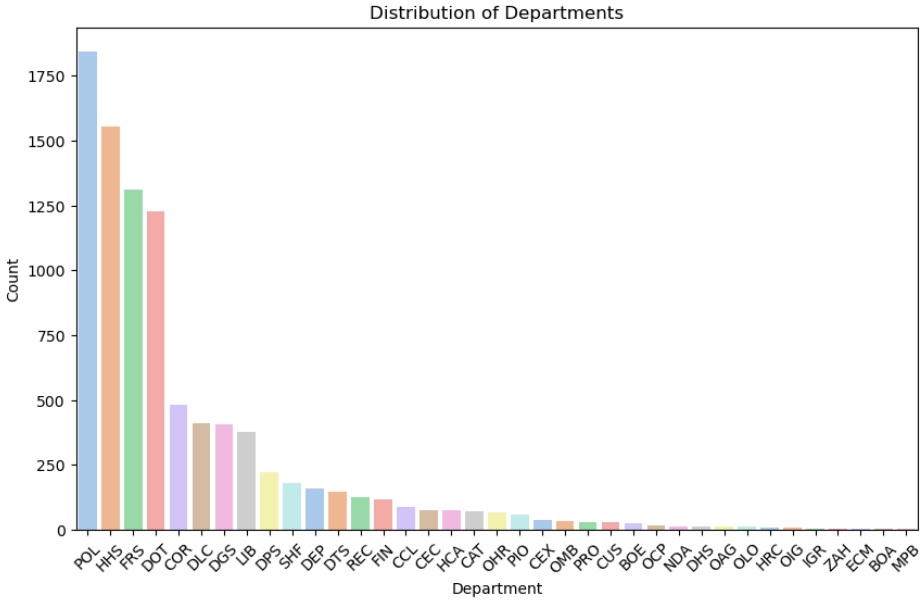

Distribution of gross pay received in 2016
Frequency
Current gross pay received in 2016
0
200
400
600
800
1000
1200
0
50000
100000
150000
200000
250000
300000

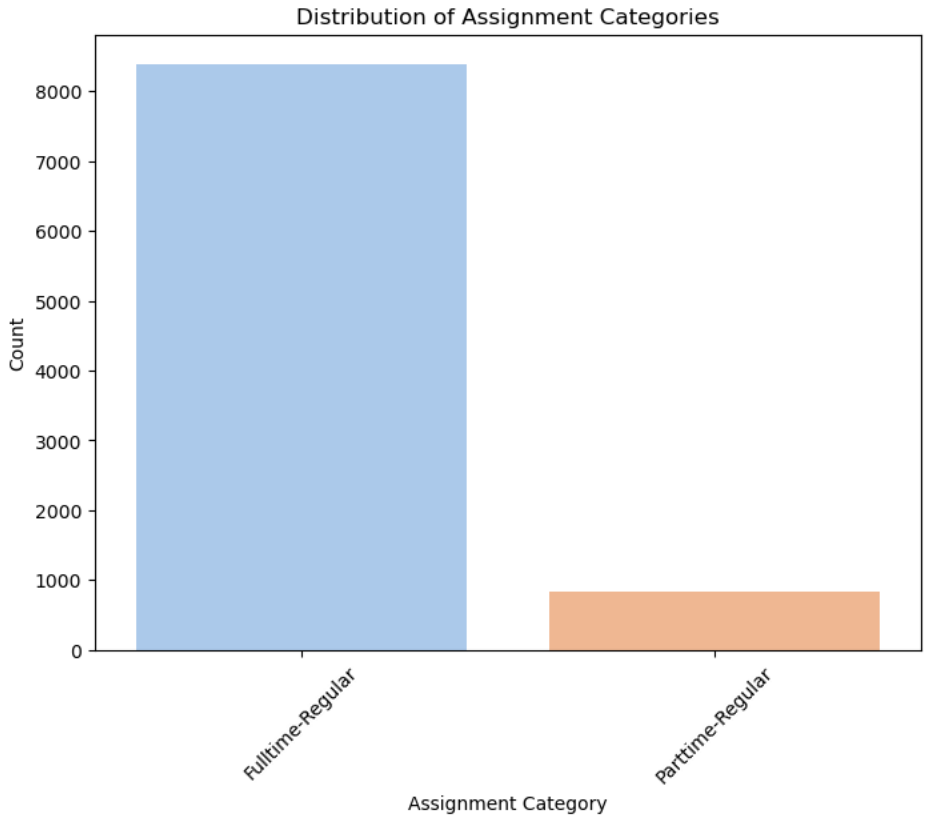

Salary Distribution by Employment Type
Salary
Employment Type
Fulltime-Regular
Parttime-Regular
0
50000
100000
150000
200000
250000
300000



2. **Histograms of Numeric Attributes**




Distribution of Departments
Count
Department
0
250
500
750
1000
1250
1500
1750
POL
HHS
FRS
DOT
COR
DLC
DGS
LIB
DPS
SHF
DEP
DTS
REC
FIN
CCL
CEC
HCA
CAT
OHR
PIO
CEX
OMB
PRO
CUS
BDE
OCP
NDA
DHS
OAG
OLO
HRC
OIG
IGR
ZAH
ECM
BOA
MPB


Distribution of Assignment Categories
Count
Assignment Category
0
1000
2000
3000
4000
5000
6000
7000
8000
Fulltime-Regular
Parttime-Regular

1.4.8 Conclusion

Both datasets offer valuable insights into weather prediction and employee compensation. By understanding their characteristics, distributions, and the significance of attributes, we can apply appropriate data preprocessing techniques and build effective predictive models.