HR Data Analytics

Data Extraction From Kaggle.com

In [1]: import pandas as pd

In [2]: #%pip install pandoc #%pip install kaggle

In [3]: import kaggle

In [4]: !kaggle datasets download hopesb/hr-analytics-dataset -f Messy_HR_Dataset_Detailed.csv

Dataset URL: https://www.kaggle.com/datasets/hopesb/hr-analytics-dataset

License(s): apache-2.0

Messy_HR_Dataset_Detailed.csv: Skipping, found more recently modified local copy (use --force to force download)

Data Cleaning & Transformation

In [5]: df = pd.read_csv('Messy_HR_Dataset_Detailed.csv')

Out[5]:

]:	Ur	nnamed: 0	FirstName	LastName	StartDate	ExitDate	Title	Supervisor	ADEmail	BusinessUnit	EmployeeStatus	•••	Satisfaction Score		Training Date	Training Program Name	Training Type	C
	0	0	Uriah	Bridges	20-Sep- 19	NaN	Production Technician I	Peter Oneill	uriah.bridges@bilearner.com	CCDR	Active		2	3	15-Jul- 23	Leadership Development	Internal	
	1	1	Paula	Small	11-Feb- 23	NaN	Production Technician I	Renee Mccormick	paula.small@bilearner.com	EW	Active		1	5	12-Sep- 22	Customer Service	External	Inc
	2	2	Edward	Buck	10-Dec- 18	NaN	Area Sales Manager	Crystal Walker	edward.buck@bilearner.com	PL	Active		2	1	13-Aug- 22	Leadership Development	External	
	3	3	Michael	Riordan	21-Jun-21	NaN	Area Sales Manager	Rebekah Wright	michael.riordan@bilearner.com	CCDR	Active		5	4	15-Dec- 22	Project Management	External	Cc
	4	4	Jasmine	Onque	29-Jun-19	NaN	Area Sales Manager	Jason Kim	jasmine.onque@bilearner.com	TNS	Active		5	3	13-Jul- 23	Technical Skills	External	
	•••													•••				
3	145	450	Oscar	Lopez	22-Oct- 18	07-Apr- 20	Network Engineer	Christian Ortega	oscar.lopez@bilearner.com	MSC	Active		3	4	04-Jun- 23	Communication Skills	External	
3	146	2164	Cruz	Arellano	17-Sep- 19	NaN	Area Sales Manager	Patricia Martinez	cruz.arellano@bilearner.com	SVG	Active		5	5	06-Oct- 22	Customer Service	Internal	Сс
3	147	2372	Makenna	Foster	20-Apr- 21	17-Sep- 22	Production Technician I	Timothy Pratt	makenna.foster@bilearner.com	CCDR	Active		1	1	28-Jul- 23	Communication Skills	Internal	Inc
3	148	2012	Cordell	Vazquez	21-Aug- 21	NaN	Production Manager	Paula Henderson	cordell.vazquez@bilearner.com	ВРС	Active		3	5	29-Oct- 22	Project Management	External	
3	149	270	Sonia	Skinner	28-Jun-20	NaN	Production Technician II	Connor Galloway	sonia.skinner@bilearner.com	PYZ	Active		4	1	24-Apr- 23	Technical Skills	Internal	

3150 rows × 39 columns

In [6]: df.drop(columns=['Unnamed: 0'], inplace=True)

Out[6]:

]:	FirstName	LastName	StartDate	ExitDate	Title	Supervisor	ADEmail	BusinessUnit	EmployeeStatus	EmployeeType	Satisfaction Score		Training Date	Training Program Name	Training Type
	0 Uriah	Bridges	20-Sep- 19	NaN	Production Technician I	Peter Oneill	uriah.bridges@bilearner.com	CCDR	Active	Contract	2	2 3	15-Jul- 23	Leadership Development	Internal
	1 Paula	Small	11-Feb- 23	NaN	Production Technician I	Renee Mccormick	paula.small@bilearner.com	EW	Active	Contract		5	12-Sep- 22	Customer Service	External
	2 Edward	Buck	10-Dec- 18	NaN	Area Sales Manager	Crystal Walker	edward.buck@bilearner.com	PL	Active	Full-Time	2	2 1	13-Aug- 22	Leadership Development	External
	3 Michael	Riordan	21-Jun-21	NaN	Area Sales Manager	Rebekah Wright	michael.riordan@bilearner.com	CCDR	Active	Contract	5	5 4	15-Dec- 22	Project Management	External
	4 Jasmine	Onque	29-Jun-19	NaN	Area Sales Manager	Jason Kim	jasmine.onque@bilearner.com	TNS	Active	Contract	5	3	13-Jul- 23	Technical Skills	External
314	5 Oscar	Lopez	22-Oct- 18	07-Apr- 20	Network Engineer	Christian Ortega	oscar.lopez@bilearner.com	MSC	Active	Part-Time	3	3 4	04-Jun- 23	Communication Skills	External
314	6 Cruz	Arellano	17-Sep- 19	NaN	Area Sales Manager	Patricia Martinez	cruz.arellano@bilearner.com	SVG	Active	Full-Time	5	5 5	06-Oct- 22	Customer Service	Internal
314	7 Makenna	Foster	20-Apr- 21	17-Sep- 22	Production Technician I	Timothy Pratt	makenna.foster@bilearner.com	CCDR	Active	Part-Time		1	28-Jul- 23	Communication Skills	Internal
314	B Cordell	Vazquez	21-Aug- 21	NaN	Production Manager	Paula Henderson	cordell.vazquez@bilearner.com	ВРС	Active	Contract	3	5	29-Oct- 22	Project Management	External
314	9 Sonia	Skinner	28-Jun-20	NaN	Production Technician II	Connor Galloway	sonia.skinner@bilearner.com	PYZ	Active	Full-Time	4	1	24-Apr- 23	Technical Skills	Internal

3150 rows × 38 columns

In [7]: date_columns = ['StartDate', 'ExitDate', 'DOB', 'Survey Date', 'Training Date'] for col in date_columns: if col in df: df[col] = pd.to_datetime(df[col], format='%d-%b-%y', errors='coerce')

#df.columns

In [8]: #df.dtypes

Cleanup Weird Characters For Column Names

```
In [9]: import pandas as pd
         import re
         def to_snake_case(name):
              name = re.sub(r'([A-Z]+)([A-Z][a-z])', r'\1_\2', name)
              name = re.sub(r'([a-z0-9])([A-Z])', r'\1_\2', name)
              name = name.replace(' ', '_')
              name = re.sub(r'[^a-zA-Z0-9_]', '', name)
              name = re.sub(r'__+', '_', name)
              return name.lower().strip('_')
         df.columns = [to_snake_case(col) for col in df.columns]
         print(df.columns)
        Index(['first_name', 'last_name', 'start_date', 'exit_date', 'title',
                'supervisor', 'ad_email', 'business_unit', 'employee_status',
                'employee_type', 'pay_zone', 'employee_classification_type',
                'termination_type', 'termination_description', 'department_type',
                'division', 'dob', 'state', 'job_function_description', 'gender_code',
                'location_code', 'race_desc', 'marital_desc', 'performance_score',
                'current_employee_rating', 'employee_id', 'survey_date',
                'engagement_score', 'satisfaction_score', 'worklife_balance_score',
                'training_date', 'training_program_name', 'training_type',
                'training_outcome', 'location', 'trainer', 'training_durationdays',
                'training_cost'],
               dtype='object')
In [10]: df
Out[10]:
                first_name last_name start_date exit_date
                                                                title supervisor
                                                                                                   ad_email business_unit employee_status employee_type ... satisfaction_score worklife_balance_score training_dat
                                                           Production
                                       2019-09-
                                                                           Peter
             0
                                                                                   uriah.bridges@bilearner.com
                                                                                                                    CCDR
                                                                                                                                                                            2
                                                                                                                                                                                                       2023-07-1
                     Uriah
                              Bridges
                                                     NaT
                                                           Technician
                                                                                                                                    Active
                                                                                                                                                 Contract ...
                                            20
                                                                          Oneill
                                                           Production
                                       2023-02-
                                                                          Renee
             1
                                                                                    paula.small@bilearner.com
                                                                                                                                                                            1
                                                                                                                                                                                                  5
                                                                                                                                                                                                       2022-09-1
                     Paula
                                Small
                                                     NaT
                                                           Technician
                                                                                                                      EW
                                                                                                                                    Active
                                                                                                                                                 Contract ...
                                             11
                                                                      Mccormick
                                       2018-12-
                                                           Area Sales
                                                                         Crystal
                                Buck
             2
                   Edward
                                                                                   edward.buck@bilearner.com
                                                                                                                      PL
                                                                                                                                    Active
                                                                                                                                                 Full-Time ...
                                                                                                                                                                            2
                                                                                                                                                                                                       2022-08-1
                                                                         Walker
                                             10
                                                            Manager
                                       2021-06-
                                                           Area Sales
                                                                        Rebekah
             3
                   Michael
                              Riordan
                                                                                 michael.riordan@bilearner.com
                                                                                                                    CCDR
                                                                                                                                    Active
                                                                                                                                                 Contract ...
                                                                                                                                                                            5
                                                                                                                                                                                                       2022-12-1
                                                            Manager
                                             21
                                                                         Wright
                                       2019-06-
                                                           Area Sales
                                                                                                                                                                            5
                                                                                                                                                                                                       2023-07-1
                   Jasmine
                              Onque
                                                                       Jason Kim jasmine.ongue@bilearner.com
                                                                                                                     TNS
                                                                                                                                    Active
                                                                                                                                                 Contract ...
                                             29
                                                            Manager
                                       2018-10- 2020-04-
                                                             Network
                                                                        Christian
          3145
                                                                                                                     MSC
                                                                                                                                    Active
                                                                                                                                                Part-Time ...
                                                                                                                                                                            3
                                                                                                                                                                                                       2023-06-0
                     Oscar
                               Lopez
                                                                                    oscar.lopez@bilearner.com
                                             22
                                                       07
                                                                         Ortega
                                                             Engineer
                                       2019-09-
                                                           Area Sales
                                                                         Patricia
                                                     NaT
          3146
                      Cruz
                             Arellano
                                                                                   cruz.arellano@bilearner.com
                                                                                                                     SVG
                                                                                                                                    Active
                                                                                                                                                 Full-Time ...
                                                                                                                                                                            5
                                                                                                                                                                                                       2022-10-0
                                             17
                                                            Manager
                                                                        Martinez
                                                           Production
                                       2021-04-
                                                2022-09-
                                                                        Timothy
                                                                                 makenna.foster@bilearner.com
                                                                                                                    CCDR
                                                                                                                                                Part-Time ...
          3147
                  Makenna
                               Foster
                                                           Technician
                                                                                                                                    Active
                                                                                                                                                                            1
                                                                                                                                                                                                       2023-07-2
                                             20
                                                       17
                                                                           Pratt
                                       2021-08-
                                                           Production
                                                                                 cordell.vazquez@bilearner.com
          3148
                                                     NaT
                                                                                                                     BPC
                                                                                                                                                                            3
                                                                                                                                                                                                  5
                                                                                                                                                                                                       2022-10-2
                   Cordell
                             Vazquez
                                                                                                                                    Active
                                                                                                                                                 Contract ...
                                                           Production
                                       2020-06-
                                                                         Connor
                     Sonia Skinner
                                                                                                                     PYZ
                                                                                                                                                                            4
                                                                                                                                                                                                     2023-04-2
          3149
                                                     NaT Technician
                                                                                   sonia.skinner@bilearner.com
                                                                                                                                    Active
                                                                                                                                                Full-Time ...
                                                                       Galloway
         3150 rows \times 38 columns
In [11]: df.drop(columns=['dob'], inplace=True)
 In [ ]: # df.to_csv('cleaned_hr_data.csv', index=False)
          Connecting To MySQL Workbench Database
In [13]: # %pip install sqlalchemy
         import sqlalchemy
In [14]: engine = sqlalchemy.create_engine('mysql+pymysql://root:Rohith18#@localhost:3306/hr_analytics')
 In [ ]: # df.to_sql('cleaned_hr_data', con=engine, if_exists='replace', index=False)
          # hr_data = pd.read_sql_table('cleaned_hr_data', engine)
         # hr_data
         Exploratory Data Analysis Using Pandas, Matplotlib & MySQL
          Question-1: Attrition Trend Analysis
          "We need to understand our employee turnover. Can you provide the monthly attrition rate for the past year, broken down by department and employee classification type?"
          Department Wise Attrition Rate (%)
In [16]: department_wise_attrition_rate = '''
          WITH exits AS (
```

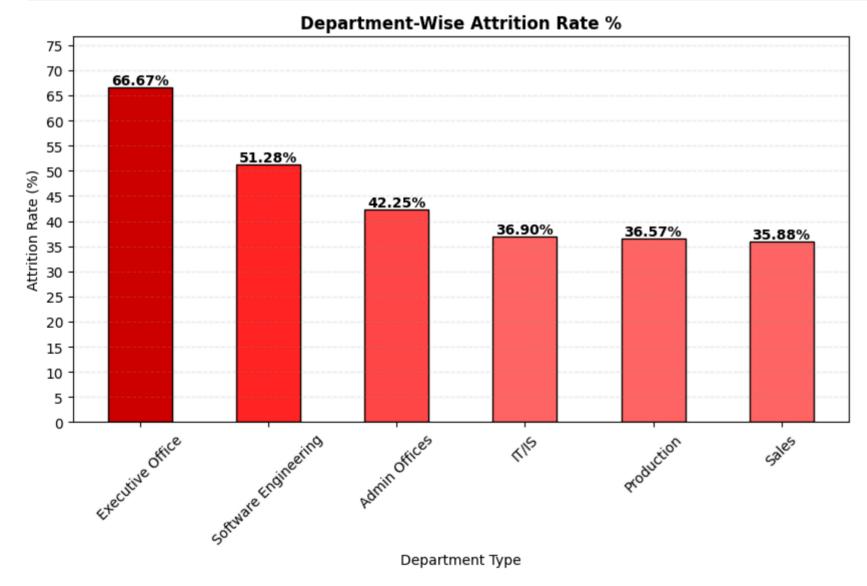
```
SELECT
        TRIM(department_type) AS department_type,
        COUNT(*) AS exits
    FROM cleaned_hr_data
    WHERE exit_date BETWEEN '2023-01-01' AND '2023-08-31'
    GROUP BY department_type
headcount_start AS (
    SELECT
       TRIM(department_type) AS department_type,
        COUNT(*) AS headcount_start
    FROM cleaned_hr_data
    WHERE start date <= '2023-01-01'
          AND (exit_date IS NULL OR exit_date >= '2023-01-01')
    GROUP BY department_type
),
headcount_end AS (
    SELECT
```

```
TRIM(department_type) AS department_type,
        COUNT(*) AS headcount_end
    FROM cleaned_hr_data
   WHERE start_date <= '2023-08-31'
         AND (exit_date IS NULL OR exit_date >= '2023-08-31')
    GROUP BY department_type
SELECT
   h_start.department_type,
   e.exits,
   h_start.headcount_start,
   h_end.headcount_end,
    ROUND((e.exits / NULLIF((h_start.headcount_start + h_end.headcount_end) / 2.0, 0)) * 100, 2) AS attrition_rate
FROM headcount_start h_start
JOIN headcount_end h_end ON h_start.department_type = h_end.department_type
LEFT JOIN exits e ON h_start.department_type = e.department_type
ORDER BY attrition_rate DESC;
departmentWiseAttritionRate = pd.read_sql_query(department_wise_attrition_rate, engine, index_col=None)
departmentWiseAttritionRate
```

Out[16]: department_type exits headcount_start headcount_end attrition_rate

0	Executive Office	4	7	5	66.67
1	Software Engineering	30	65	52	51.28
2	Admin Offices	15	37	34	42.25
3	IT/IS	88	259	218	36.90
4	Production	422	1247	1061	36.57
5	Sales	68	205	174	35.88

```
In [33]: import matplotlib.pyplot as plt
         plt.figure(figsize=(10, 5))
         colors = departmentWiseAttritionRate['attrition_rate'].apply(
             lambda i: '#D00000' if i > 60 else '#FF2525' if i > 50 else '#FF4747' if i > 40 else '#FF6565'
         bars = plt.bar(departmentWiseAttritionRate['department_type'],
                        departmentWiseAttritionRate['attrition_rate'],
                        color=colors, width=0.5, edgecolor="black")
         max_attrition = departmentWiseAttritionRate['attrition_rate'].max()
         plt.ylim(0, max_attrition + 10)
         plt.yticks(range(0, int(max_attrition) + 10, 5))
         for bar in bars:
             plt.text(bar.get_x() + bar.get_width()/2,
                      bar.get_height(),
                      f'{bar.get_height():.2f}%',
                      ha='center', va='bottom', fontsize=10, color='black', fontdict={'fontweight': 'semibold'})
         plt.xlabel('Department Type')
         plt.ylabel('Attrition Rate (%)')
         plt.title('Department-Wise Attrition Rate %', fontdict={'fontsize': 12, 'fontweight': 'semibold'})
         plt.xticks(rotation=45)
         plt.grid(axis='y', linestyle='--', alpha=0.2)
         plt.show()
```



Employee Classification Type Wise Attrition Rate (%)

```
In [18]: classification_type_wise_attrition_rate = '''
         WITH exits AS (
             SELECT
                 TRIM(employee_classification_type) AS classification_type,
                 COUNT(*) AS exits
             FROM cleaned_hr_data
             WHERE exit_date BETWEEN '2023-01-01' AND '2023-08-31'
             GROUP BY classification_type
         ),
         headcount_start AS (
             SELECT
                 TRIM(employee_classification_type) AS classification_type,
                 COUNT(*) AS headcount_start
             FROM cleaned_hr_data
             WHERE start_date <= '2023-01-01'
                   AND (exit date IS NULL OR exit date >= '2023-01-01')
             GROUP BY classification type
         headcount_end AS (
             SELECT
                 TRIM(employee_classification_type) AS classification_type,
```

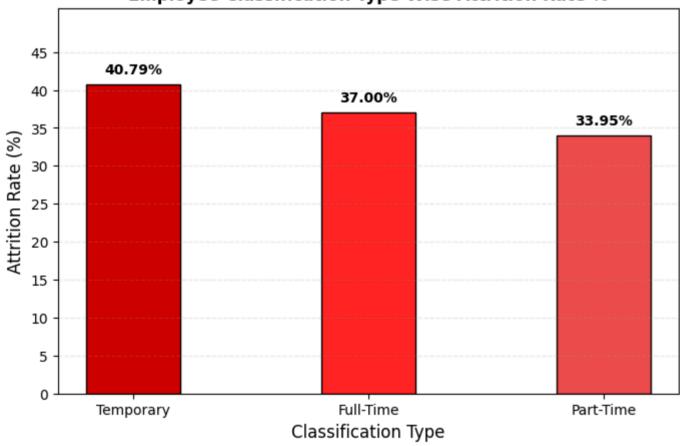
```
COUNT(*) AS headcount_end
    FROM cleaned_hr_data
    WHERE start_date <= '2023-08-31'
          AND (exit_date IS NULL OR exit_date >= '2023-08-31')
    GROUP BY classification_type
SELECT
   h_start.classification_type,
   e.exits,
   h_start.headcount_start,
   h_end.headcount_end,
   ROUND((e.exits / NULLIF((h_start.headcount_start + h_end.headcount_end) / 2.0, 0)) * 100, 2) AS attrition_rate
FROM headcount_start h_start
JOIN headcount_end h_end ON h_start.classification_type = h_end.classification_type
LEFT JOIN exits e ON h_start.classification_type = e.classification_type
ORDER BY attrition_rate DESC;
classificationTypeWiseAttritionRate = pd.read_sql_query(classification_type_wise_attrition_rate, engine)
classificationTypeWiseAttritionRate
```

Out[18]: classification_type exits headcount_start headcount_end attrition_rate

0	Temporary	227	609	504	40.79
1	Full-Time	217	636	537	37.00
2	Part-Time	183	575	503	33.95

```
In [32]: import matplotlib.pyplot as plt
         plt.figure(figsize=(8, 5))
         colors = classificationTypeWiseAttritionRate['attrition_rate'].apply(
             lambda i: '#D00000' if i > 40 else '#FF2525' if i > 35 else '#EE4E4E'
         bars = plt.bar(classificationTypeWiseAttritionRate['classification_type'],
                        classificationTypeWiseAttritionRate['attrition_rate'],
                        color=colors, width=0.4, edgecolor="black")
         max_attrition = classificationTypeWiseAttritionRate['attrition_rate'].max()
         plt.ylim(0, max_attrition + 10)
         plt.yticks(range(0, int(max_attrition) + 10, 5))
         for bar in bars:
             plt.text(bar.get_x() + bar.get_width() / 2,
                      bar.get_height() + 1,
                      f'{bar.get_height():.2f}%',
                      ha='center', va='bottom', fontsize=10, color='black', fontweight='semibold')
         plt.xlabel('Classification Type', fontsize=12)
         plt.ylabel('Attrition Rate (%)', fontsize=12)
         plt.title('Employee Classification Type-Wise Attrition Rate %', fontsize=12, fontweight='semibold')
         plt.xticks(rotation=0)
         plt.grid(axis='y', linestyle='--', alpha=0.2)
         plt.show()
```

Employee Classification Type-Wise Attrition Rate %



Monthly Attrition Rate (%)

```
In [20]: monthly_attrition_rate = """
         WITH monthly_exits AS (
             SELECT
                  DATE_FORMAT(exit_date, '%%Y-%m') AS month_year,
                  COUNT(*) AS exits
              FROM hr_analytics.cleaned_hr_data
              WHERE exit_date BETWEEN '2023-01-01' AND '2023-08-31'
             GROUP BY month_year
         headcount_start AS (
             SELECT
                  DATE_FORMAT(start_date, '%Y-%m') AS month_year,
                  COUNT(*) AS start_count
              FROM hr_analytics.cleaned_hr_data
             WHERE start_date <= '2023-08-31'
             GROUP BY month_year
         headcount_end AS (
             SELECT
                  DATE_FORMAT(start_date, '%%Y-%m') AS month_year,
                  COUNT(*) AS end_count
             FROM hr_analytics.cleaned_hr_data
              WHERE start_date <= '2023-08-31'</pre>
             AND (exit_date IS NULL OR exit_date > LAST_DAY(start_date))
              GROUP BY month_year
         SELECT
              months.month_year,
              COALESCE(me.exits, 0) AS exits,
              COALESCE(hs.start_count, 0) AS headcount_start,
              COALESCE(he.end_count, 0) AS headcount_end,
              ROUND(((COALESCE(hs.start_count, 0) + COALESCE(he.end_count, 0)) / 2), 2) AS avg_headcount,
              ROUND((COALESCE(me.exits, 0) \ / \ NULLIF(((COALESCE(hs.start\_count, 0) + COALESCE(he.end\_count, 0)) \ / \ 2), \ 0)) \ * \ 100, \ 2) \ AS \ attrition\_rate
```

```
FROM (

SELECT month_year FROM headcount_start WHERE month_year BETWEEN '2023-01' AND '2023-08'
UNION

SELECT month_year FROM headcount_end WHERE month_year BETWEEN '2023-01' AND '2023-08'
UNION

SELECT month_year FROM monthly_exits WHERE month_year BETWEEN '2023-01' AND '2023-08'
) AS months

LEFT JOIN headcount_start hs ON months.month_year = hs.month_year

LEFT JOIN headcount_end he ON months.month_year = he.month_year

LEFT JOIN monthly_exits me ON months.month_year = me.month_year

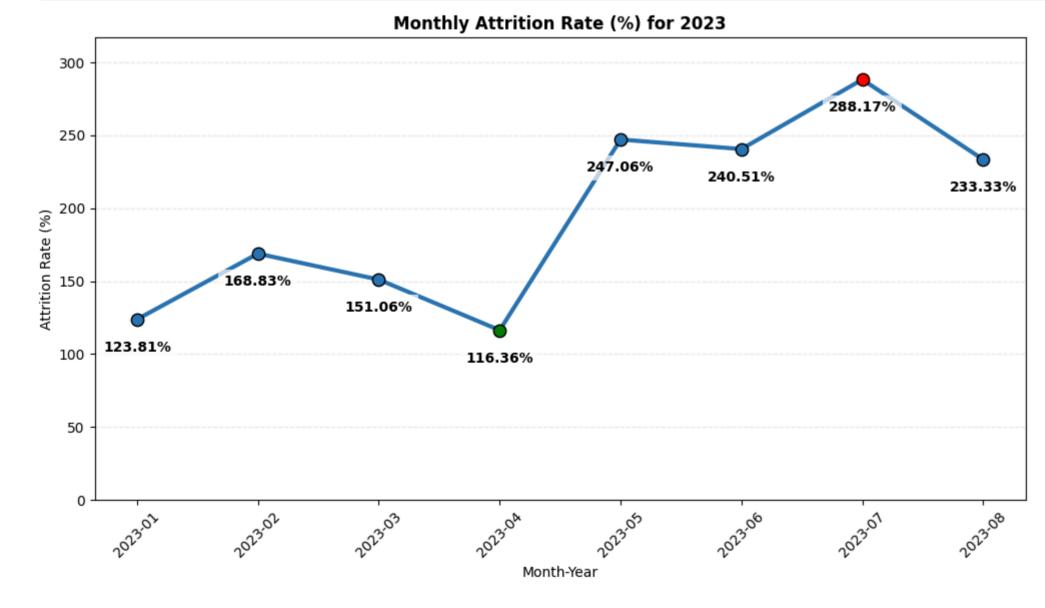
ORDER BY months.month_year;

"""

monthlyAttritionRate = pd.read_sql_query(monthly_attrition_rate, engine)
monthlyAttritionRate
```

Out[20]: month_year exits headcount_start headcount_end avg_headcount attrition_rate 0 2023-01 65 53 52 52.5 123.81 65 38 38.5 2023-02 39 168.83 43 2 2023-03 71 51 47.0 151.06 3 2023-04 64 55 55 55.0 116.36 45 40 4 2023-05 105 42.5 247.06 40 39 39.5 2023-06 95 240.51 57 6 2023-07 134 36 46.5 288.17 7 14 10 2023-08 28 12.0 233.33

```
In [21]: import matplotlib.pyplot as plt
         max_value = max(monthlyAttritionRate['attrition_rate'])
         min_value = min(monthlyAttritionRate['attrition_rate'])
         plt.figure(figsize=(12, 6))
         plt.plot(monthlyAttritionRate['month_year'],
                  monthlyAttritionRate['attrition_rate'],
                  linestyle='-', color='#2973B2', linewidth=3)
         for i, value in enumerate(monthlyAttritionRate['attrition_rate']):
             if value == max_value:
                 color = 'red'
             elif value == min_value:
                 color = 'green'
             else:
                 color = '#2973B2'
             plt.scatter(monthlyAttritionRate['month_year'][i], value, color=color, s=80, edgecolors='black', zorder=3)
             plt.text(monthlyAttritionRate['month_year'][i],
                      value - (max_value * 0.05),
                      f'{value:.2f}%',
                      ha='center', va='top', fontsize=10, color='black', fontdict={'fontweight': 'semibold'},
                      bbox=dict(facecolor='white', edgecolor='none', alpha=0.7))
         plt.xlabel('Month-Year')
         plt.ylabel('Attrition Rate (%)')
         plt.title('Monthly Attrition Rate (%) for 2023', fontdict={'fontsize': 12, 'fontweight': 'semibold'})
         plt.xticks(rotation=45)
         plt.ylim(0, max_value * 1.1)
         plt.grid(axis='y', linestyle='--', alpha=0.2)
         plt.show()
```

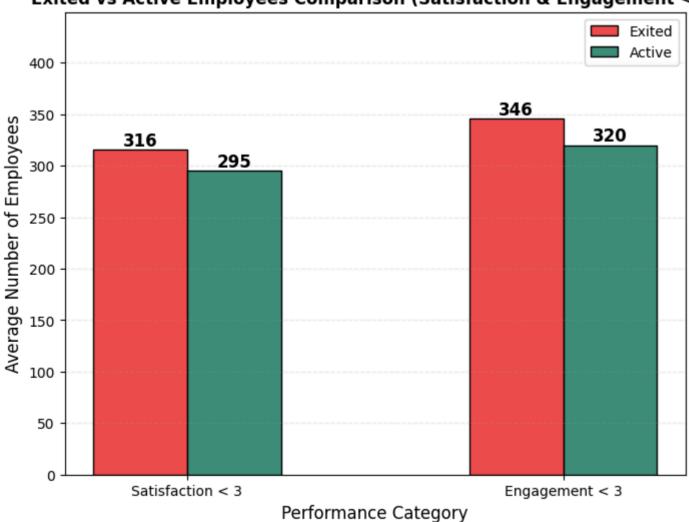


Question-2: Performance & Exit Insights

"I suspect that employees with low satisfaction scores are more likely to leave. Can you analyze whether employees who have exited had lower satisfaction and engagement scores compared to those still active?

```
engagement_score as active_engagement_score, COUNT(*) AS total_employees
         FROM
              cleaned hr data
         WHERE exit_date IS NULL AND engagement_score < 3</pre>
         GROUP BY engagement_score
         ORDER BY engagement_score),
         active_satisfaction_score as (
              satisfaction_score as active_satisfaction_score, COUNT(*) AS total_employees
         FROM
              cleaned_hr_data
         WHERE exit_date IS NULL AND satisfaction_score < 3
         GROUP BY satisfaction_score
         ORDER BY satisfaction_score)
         SELECT
              ROUND(AVG(s.total_employees), 2) AS avg_exited_employees_with_below_3_satisfaction_score,
              ROUND(AVG(e.total_employees), 2) AS avg_exited_employees_with_below_3_engagement_score,
             ROUND(AVG(active_s.total_employees), 2) AS avg_active_employees_with_below_3_satisfaction_score,
              ROUND(AVG(active_e.total_employees), 2) AS avg_active_employees_with_below_3_engagement_score
              active_satisfaction_score active_s,
              active_engagement_score active_e,
              exited_satisfaction_score s,
              exited_engagement_score e;
         activeVsExitedEmployeesPerformance = pd.read_sql_query(active_vs_exited_employees_performance, engine)
         activeVsExitedEmployeesPerformance
Out[22]:
            avg_exited_employees_with_below_3_satisfaction_score avg_exited_employees_with_below_3_engagement_score avg_active_employees_with_below_3_satisfaction_score avg_active_employees_with_below_3_engageme
          0
                                                        315.5
                                                                                                           345.5
                                                                                                                                                             295.0
In [31]: import pandas as pd
         import matplotlib.pyplot as plt
         import numpy as np
         df = activeVsExitedEmployeesPerformance
         df.columns = [
             "Exited (Satisfaction < 3)",</pre>
              "Exited (Engagement < 3)",</pre>
              "Active (Satisfaction < 3)",
              "Active (Engagement < 3)"
         categories = ["Satisfaction < 3", "Engagement < 3"]</pre>
         exited_values = [df.iloc[0]["Exited (Satisfaction < 3)"], df.iloc[0]["Exited (Engagement < 3)"]]</pre>
         active_values = [df.iloc[0]["Active (Satisfaction < 3)"], df.iloc[0]["Active (Engagement < 3)"]]</pre>
         x = np.arange(len(categories))
         width = 0.25
         fig, ax = plt.subplots(figsize=(8, 6))
         bars1 = ax.bar(x - width/2, exited_values, width, color="#EE4E4E", label="Exited", edgecolor="black")
         bars2 = ax.bar(x + width/2, active_values, width, color="#3D8D7A", label="Active", edgecolor="black")
         for bar in bars1 + bars2:
              height = bar.get_height()
              ax.text(bar.get_x() + bar.get_width()/2, height, f'{height:.0f}',
                      ha='center', va='bottom', fontsize=12, fontweight='bold')
         ax.set_xlabel("Performance Category", fontsize=12)
         ax.set_ylabel("Average Number of Employees", fontsize=12)
         ax.set_title("Exited vs Active Employees Comparison (Satisfaction & Engagement < 3)", fontsize=12, fontweight='semibold')</pre>
         ax.set_ylim(0, max(max(exited_values), max(active_values)) * 1.3)
         ax.yaxis.set_major_locator(plt.MaxNLocator(integer=True))
         ax.yaxis.grid(True, linestyle="--", alpha=0.2)
         ax.set_xticks(x)
         ax.set_xticklabels(categories)
         ax.legend()
         plt.show()
```

Exited vs Active Employees Comparison (Satisfaction & Engagement < 3)



```
Question-3: Training Effectiveness on Performance
```

active_engagement_score as (SELECT

"We've been investing heavily in training programs. Can you analyze whether employees who completed training in the past year have a lower attrition rate compared to those who didn't?"

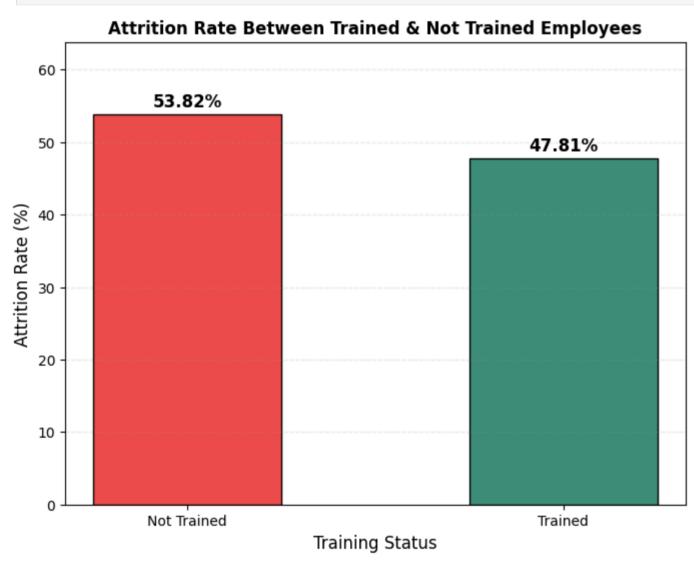
```
In [24]: attrition_rate_between_trained_and_untrained_employees = '''
WITH trained_employees AS (
    SELECT employee_id
    FROM hr_analytics.cleaned_hr_data
    WHERE training_date BETWEEN '2023-08-01' - INTERVAL 11 MONTH AND '2023-08-31' AND training_outcome IN ('Completed', 'Passed')
),
```

```
attrition_data AS (
    SELECT
       CASE
            WHEN e.employee_id IN (SELECT employee_id FROM trained_employees) THEN 'Trained'
            ELSE 'Not Trained'
       END AS training_status,
       COUNT(*) AS total_employees,
       SUM(CASE WHEN e.exit_date IS NOT NULL THEN 1 ELSE 0 END) AS exited_employees
    FROM hr analytics.cleaned hr data e
   GROUP BY training_status
SELECT
    training_status,
    total_employees,
    exited_employees,
    ROUND((exited_employees / NULLIF(total_employees, 0)) * 100, 2) AS attrition_rate
FROM attrition_data;
attritionRateBetweenTrainedAndUntrainedEmployees = pd.read_sql_query(attrition_rate_between_trained_and_untrained_employees, engine)
attritionRateBetweenTrainedAndUntrainedEmployees
```

Out[24]: training_status total_employees exited_employees attrition_rate

0	Not Trained	1663	895.0	53.82
1	Trained	1487	711.0	47.81

```
In [25]: import matplotlib.pyplot as plt
         fig, ax = plt.subplots(figsize=(8, 6))
         ax.bar(
             attritionRateBetweenTrainedAndUntrainedEmployees["training_status"],
             attritionRateBetweenTrainedAndUntrainedEmployees["attrition_rate"],
             color=["#EE4E4E", "#3D8D7A"], width=0.5, edgecolor="black"
         ax.set xlabel("Training Status", fontsize=12)
         ax.set ylabel("Attrition Rate (%)", fontsize=12)
         ax.set_title("Attrition Rate Between Trained & Not Trained Employees", fontsize=12, fontweight='semibold')
         max_attrition = attritionRateBetweenTrainedAndUntrainedEmployees["attrition_rate"].max()
         ax.set_ylim(0, max_attrition + 10)
         ax.set_yticks(range(0, int(max_attrition) + 10, 10))
         ax.set_xticks(range(len(attritionRateBetweenTrainedAndUntrainedEmployees["training_status"])))
         ax.set_xticklabels(attritionRateBetweenTrainedAndUntrainedEmployees["training_status"], rotation=0)
         for i, v in enumerate(attritionRateBetweenTrainedAndUntrainedEmployees["attrition_rate"]):
             ax.text(i, v + 1, f"{v:.2f}%", ha='center', fontsize=12, fontweight='semibold')
         ax.yaxis.grid(True, linestyle="--", alpha=0.2)
         plt.show()
```



Question-4: Diversity & Inclusion Hiring Metrics

"I need to prepare a report for leadership on our diversity hiring efforts. Can you provide a breakdown of new hires in the past two years by gender and job function?"

2

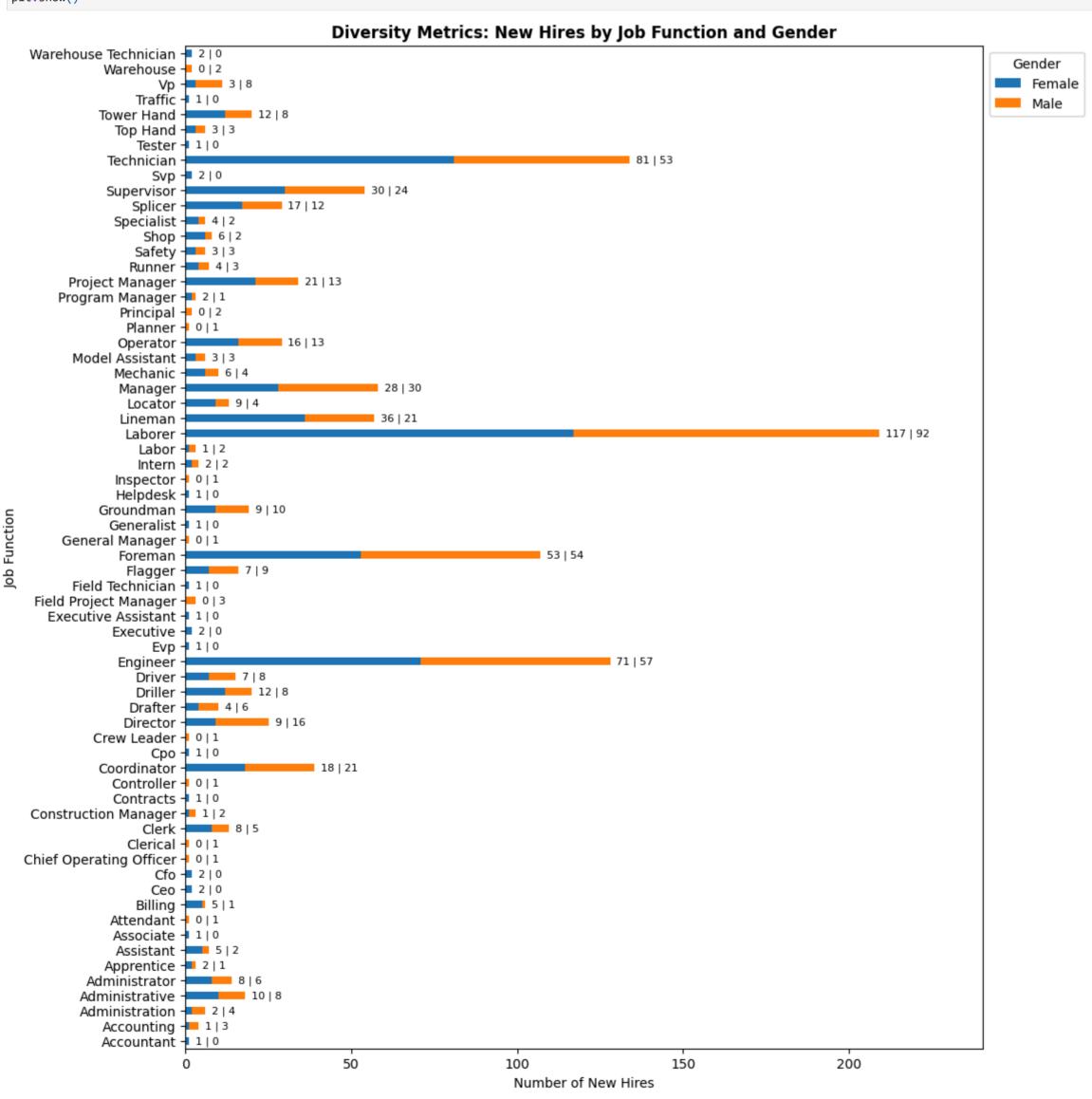
Female

105 rows × 3 columns

Warehouse Technician

104

```
In [27]: #%pip install mplcursors
         # %pip install seaborn
In [28]: import matplotlib.pyplot as plt
         pivot_df = diversityMetrices.pivot(index="job_function_description", columns="gender_code", values="new_hires").fillna(0)
         fig, ax = plt.subplots(figsize=(14, len(pivot_df) * 0.2))
         colors = ["#1f77b4", "#ff7f0e"]
         pivot_df.plot(kind="barh", stacked=True, color=colors[:len(pivot_df.columns)], ax=ax)
         for index, (job_function, row) in enumerate(pivot_df.iterrows()):
             gender_counts = [int(row.get(gender, 0)) for gender in pivot_df.columns]
             total = sum(gender_counts)
             if total > 0:
                 label_text = " | ".join(str(count) for count in gender_counts)
                 ax.text(total + 2, index, label_text, va='center', fontsize=8, fontweight='normal')
         ax.set_xlabel("Number of New Hires")
         ax.set_ylabel("Job Function")
         ax.set_title("Diversity Metrics: New Hires by Job Function and Gender", fontdict={'fontsize': 12, 'fontweight': 'semibold'})
         ax.legend(title="Gender", bbox_to_anchor=(1, 1), loc='upper left')
         plt.subplots_adjust(left=0.2, right=0.8, top=0.9, bottom=0.1)
         plt.gca().margins(x=0.15, y=0.01)
         plt.show()
```



"Some employees have raised concerns about work-life balance. Can you analyze whether employees with low work-life balance scores are more likely to leave or have lower performance ratings?"

```
In [29]: work_life_balance_vs_attrition_rate = '''
         WITH exit_employees_ratings AS (SELECT
             performance_score,
             COUNT(*) AS total_employees,
             COUNT(exit_date) AS exits,
             ROUND(AVG(worklife_balance_score), 2) AS avg_work_life_balance_rating
             cleaned_hr_data
         GROUP BY performance_score
         SELECT
             performance_score,
             total employees,
             avg_work_life_balance_rating,
             ROUND((exits / total_employees) * 100, 2) AS attrition_rate
             exit_employees_ratings
         ORDER BY attrition_rate DESC;
         workLifeBalanceVsAttritionRate = pd.read_sql_query(work_life_balance_vs_attrition_rate, engine)
         workLifeBalanceVsAttritionRate
```

Out[29]: performance_score total_employees exits avg_work_life_balance_rating attrition_rate 0 51 2.99 52.58 1 Needs Improvement 2.91 51.09 184 94 2 **Fully Meets** 3.00 2486 1268 51.01 3 383 193 Exceeds 2.95 50.39

```
In [30]: import matplotlib.pyplot as plt
         import numpy as np
         x = workLifeBalanceVsAttritionRate["performance_score"]
         y1 = workLifeBalanceVsAttritionRate["attrition_rate"]
         y2 = workLifeBalanceVsAttritionRate["avg_work_life_balance_rating"]
         fig, ax1 = plt.subplots(figsize=(12, 6))
         bars = ax1.bar(x, y1, color="#EE4E4E", width=0.3, edgecolor="black", label="Attrition Rate (%)")
         ax2 = ax1.twinx()
         ax2.plot(x, y2, color="#FBA518", marker="o", linestyle="--", linewidth=2, markersize=8, label="Work-Life Balance Score")
         ax1.set_xlabel("Performance Score", fontsize=12)
         ax1.set_ylabel("Attrition Rate (%)", fontsize=12)
         ax2.set_ylabel("Avg Work-Life Balance Score", fontsize=12)
         plt.title("Performance Score vs Attrition Rate & Work-Life Balance For Exited Employees", fontsize=14, fontweight="bold")
         ax1.set_xticks(range(len(x)))
         ax1.set_xticklabels(x, rotation=0, ha='center', fontsize=12)
         ax1.set_ylim(0, max(y1) + 15)
         ax2.set_ylim(0, max(y2) + 2)
         for bar in bars:
             height = bar.get_height()
             ax1.text(bar.get_x() + bar.get_width()/2, height + 1, f"{height:.2f}%", ha="center", fontsize=12, fontweight="bold")
         for i, txt in enumerate(y2):
             ax2.text(i, y2[i] - 0.4, f"{txt:.2f}", ha="center", fontsize=12, fontweight="bold", color="white")
         ax1.grid(axis="y", linestyle="--", alpha=0.2)
         ax1.legend(loc="upper left")
         ax2.legend(loc="upper right")
         plt.show()
```

