HHRR Employee

CONTENIDO

INTRODUCCIÓN.

Este es un conjunto de datos creado por los científicos de datos de IBM para analizar los factores que provocan el abandono de los empleados.
La elección de este, entre otros dos, nos representó un desafío interesante, ya que consideramos que el desgaste (Attrition) es un problema que afecta a todas las empresas, independientemente de su ubicación geográfica, sector y tamaño. El desgaste de los empleados genera costes significativos, incluidos los derivados de la interrupción del negocio, la contratación de nuevo personal y su formación. Se buscará conocer los factores que impulsan y minimizan la rotación de personal.

OBJETIVO.

El objetivo del modelo es predecir la probabilidad de la variable objetivo, “desgaste” (Attrition) anticipando el abandono de los empleados, en función de diversas características de estos.
Se analizará cuales son los factores que generan un mayor desgaste laboral, que llevan a una posible deserción de los empleados de la organización, estudiando los motivos (voluntario o involuntario), incluida la dimisión, el despido, o la jubilación.

META.

Clasificar
Predecir si un empleado continúa o no en la empresa.
Extraer información sobre el rendimiento de los empleados, así como la retención de este.

PLANTEAMIENTO DEL PROBLEMA.

Este conjunto de datos ficticios nos brinda la oportunidad de automatizar el sistema de contratación y despido de empleados de una organización.
Predeciremos el desgaste de los empleados (si un empleado renunciará o no), basándonos en las siguientes variables:

0) Age (Edad)
1) Attrition (Se siente desgastado);

2) BusinessTravel (Viajes de negocio);
3) DailyRate (Importe diario);
4) Department (Departamento);
5) DistanceFromHome (Distancia desde casa);
6) Education (Educación);
7) EducationField (Área educativa);
8) EmployeeCount (Número de empleados);
9) EmployeeNumber (Número de empleado);
10) EnvironmentSatisfaction (Satisfacción del ambiente laboral);
11) Gender (Genero);
12) HourlyRate (Importe por Hs.);
13) JobInvolvement (Participación);
14) JobLevel (Nivel);
15) JobRole (Rol);
16) JobSatisfaction (Satisfacción laboral);
17) MaritalStatus (Estado civil);
18) MonthlyIncome (Ingresos mensuales);
19) MonthlyRate (Importe mensual);
20) NumCompaniesWorked (Nro empresas trabajadas);
21) Over18 (Mayores de 18);
22) OverTime (Horas extras);
23) PercentSalaryHike (% aumento salarial);
24) PerformanceRating (Clasificación de rendimiento);
25) RelationshipSatisfaction (Relación satisfacción)

26) StandardHours (Horas comunes);
27) StockOptionLevel (Nivel de acciones);
28) TotalWorkingYears (Total años trabajados);
29) TrainingTimesLastYear (Tiempo de entrenamiento del año pasado);
30) WorkLifeBalance (Equilibrio vida-trabajo);
31) YearsAtCompany (Años en la empresa);
32) YearsInCurrentRole (Años en el rol actual);
33) YearsSinceLastPromotion (Años desde la última promoción);
34) YearsWithCurrManager (Años como gerente).

ALGUNAS PREGUNTAS QUE BUSCAREMOS RESPONDER.

- ¿Cuáles son los factores que influyen en la tasa de abandono?
- ¿Cuál es la tasa de abandono? ¿Es alta o baja?
- ¿La distancia desde la casa al trabajo es un factor determinante para cambiar de trabajo?
- ¿Cuántas personas hay según los niveles de trabajo?
- ¿Qué nivel de estudio poseen?
- ¿Cuál es el grado de satisfacción con la empresa?
- ¿Cuál es el rango de salario en la empresa?
- ¿Cuánto más antigüedad en la empresa, mayor es el salario?
- ¿A mayor edad se incrementa el salario?
- ¿El salario aumenta respecto al nivel que tiene cada cargo?
- ¿Cómo es en general el nivel de balance entre el trabajo y la vida de los empleados?

- ¿Qué relación tienen los empleados con un mismo jefe con respecto al desgaste?
- ¿Qué recomendaciones se pueden dar a la dirección basándose en los análisis?

TABLA DE VERSIONADOS.

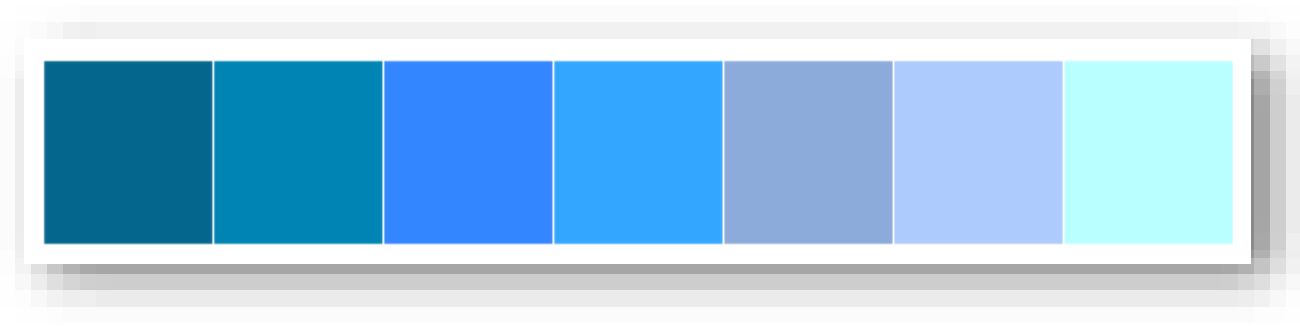
Elección de potenciales DataSets.	Habilitada: 22/10/2022
Práctica integradora: Visualizaciones en Python.	Habilitada: 05/11/2022
Estructurando un Proyecto de DS-Parte I	Habilitada: 19/11/2022
Estructurando un Proyecto de DS-Parte II	Habilitada: 03/12/2022
Estructurando un Proyecto de DS-Parte III	Habilitada: 17/12/2022
Evaluando modelos ML	Habilitada: 17/12/2022
Primera Pre-Entrega del Proyecto Final	Habilitada: 07/01/2023
Data Wrangling	Habilitada: 11/02/2023
Data Storytelling	Habilitada: 25/02/2023
Obtención de Insights	Habilitada: 11/03/2023
Segunda pre-entrega de Proyecto Final	Habilitada: 18/03/2023
Evaluando modelos de Machine Learning	Habilitada: 08/04/2023
Ingeniería de atributos y selección de variables	Habilitada: 15/04/2023
Proyecto Final	Habilitado:13/05/2023

DESCRIPCIÓN DE IBM HHRR ATTRITION.

IBM HHRR Attrition es un conjunto de datos con más de 30 variables características categóricas y discretas con datos numéricos y de texto, con 1470 registros.
Es un dataset público, y ha sido descargado del sitio Kaggle.

PALETA DE COLORES EN GRÁFICOS.

Se eligió una gama de colores dentro del azul, ya que la paleta utilizada por IBM se encuentra en esa gama, y también para una visualización más prolija y atractiva.



ESTADÍSTICA DESCRIPTIVA DE LAS COLUMNAS NUMÉRICAS.

	count	mean	std	min	25%	50%	75%	max
Age	2149.0	37.05	9.25	18.0	30.0	36.0	43.0	60.0
DailyRate	2149.0	806.53	405.70	102.0	465.0	809.0	1158.0	1499.0
DistanceFromHome	2149.0	9.17	8.10	1.0	2.0	7.0	14.0	29.0
Education	2149.0	2.90	1.02	1.0	2.0	3.0	4.0	5.0
EmployeeCount	2149.0	1.00	0.00	1.0	1.0	1.0	1.0	1.0
EmployeeNumber	2149.0	1075.00	620.51	1.0	538.0	1075.0	1612.0	2149.0
EnvironmentSatisfaction	2149.0	2.72	1.09	1.0	2.0	3.0	4.0	4.0
HourlyRate	2149.0	66.16	20.35	30.0	48.0	66.0	84.0	100.0
JobInvolvement	2149.0	2.73	0.71	1.0	2.0	3.0	3.0	4.0
JobLevel	2149.0	2.08	1.13	1.0	1.0	2.0	3.0	5.0
JobSatisfaction	2149.0	2.75	1.10	1.0	2.0	3.0	4.0	4.0
MonthlyIncome	2149.0	6523.07	4753.57	1009.0	2875.0	4907.0	8446.0	19999.0
MonthlyRate	2149.0	14247.12	7052.85	2094.0	8202.0	14074.0	20338.0	26999.0
NumCompaniesWorked	2149.0	2.68	2.50	0.0	1.0	2.0	4.0	9.0
PercentSalaryHike	2149.0	15.28	3.67	11.0	12.0	14.0	18.0	25.0
PerformanceRating	2149.0	3.16	0.36	3.0	3.0	3.0	3.0	4.0

RelationshipSatisfaction	2149.0	2.71	1.09	1.0	2.0	3.0	4.0	4.0
StandardHours	2149.0	80.00	0.00	80.0	80.0	80.0	80.0	80.0
StockOptionLevel	2149.0	0.79	0.85	0.0	0.0	1.0	1.0	3.0
TotalWorkingYears	2149.0	11.28	7.81	0.0	6.0	10.0	15.0	40.0
TrainingTimesLastYear	2149.0	2.78	1.29	0.0	2.0	3.0	3.0	6.0
WorkLifeBalance	2149.0	2.75	0.70	1.0	2.0	3.0	3.0	4.0
YearsAtCompany	2149.0	7.03	6.18	0.0	3.0	5.0	9.0	40.0
YearsInCurrentRole	2149.0	4.25	3.67	0.0	2.0	3.0	7.0	18.0
YearsSinceLastPromotion	2149.0	2.18	3.22	0.0	0.0	1.0	3.0	15.0
YearsWithCurrManager	2149.0	4.13	3.60	0.0	2.0	3.0	7.0	17.0

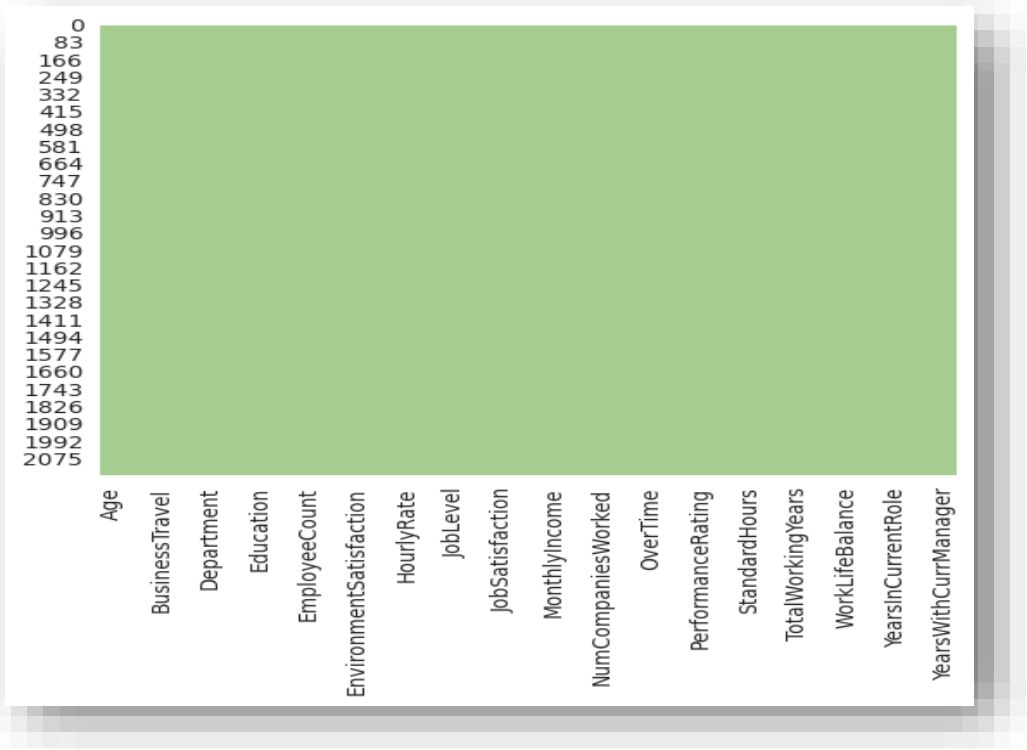
Las filas son los atributos (nombre de los campos) y las columnas nos da la cantidad de registros, los valores máximos y mínimos, los cuartiles, la media y la desviación estándar.

ESTADÍSTICAS DE LAS COLUMNAS DE CATEGÓRICAS.

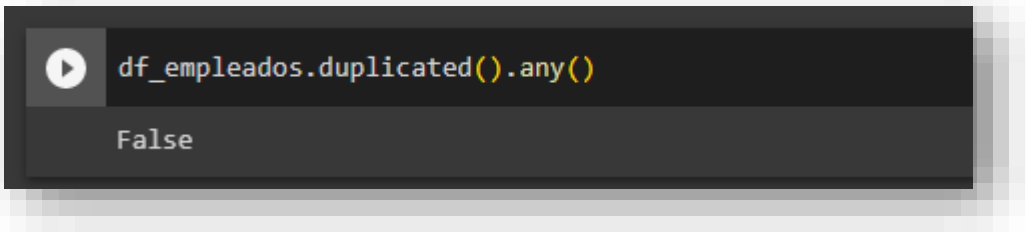
	Attrition	BusinessTravel	Department	EducationField	Gender	JobRole	MaritalStatus	Over18	OverTime
count	2149	2149	2149	2149	2149	2149	2149	2149	2149
unique	2	3	3	6	2	9	3	1	2
top	No	Travel_Rarely	Research & Development	Life Sciences	Male	Sales Executive	Married	Y	No
freq	1795	1509	1432	894	1285	461	973	2149	1529
missing_rate	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

ANÁLISIS DE LOS DATOS.

LOS DATOS NO CONTIENEN VALORES NULOS.



LOS DATOS NO SE ENCUENTRAN DUPLICADOS.



TIPO DE DATOS INICIAL.

#	Column:	Non-Null Count	Dtype
0	Age	2149 non	null int64
1	Attrition	2149 non	null object
2	BusinessTravel	2149 non	null object
3	DailyRate	2149 non	null int64
4	Department	2149 non	null object
5	DistanceFromHome	2149 non	null int64

6	Education	2149 non	null	int64
7	EducationField	2149 non	null	object
8	EmployeeCount	2149 non	null	int64
9	EmployeeNumber	2149 non	null	int64
10	EnvironmentSatisfaction	2149 non	null	int64
11	Gender	2149 non	null	object
12	HourlyRate	2149 non	null	int64
13	JobInvolvement	2149 non	null	int64
14	JobLevel	2149 non	null	int64
15	JobRole	2149 non	null	object
16	JobSatisfaction	2149 non	null	int64
17	MaritalStatus	2149 non	null	object
18	MonthlyIncome	2149 non	null	int64
19	MonthlyRate	2149 non	null	int64
20	NumCompaniesWorked	2149 non	null	int64
21	Over18	2149 non	null	object
22	OverTime	2149 non	null	object
23	PercentSalaryHike	2149 non	null	int64
24	PerformanceRating	2149 non	null	int64
25	RelationshipSatisfaction	2149 non	null	int64
26	StandardHours	2149 non	null	int64
27	StockOptionLevel	2149 non	null	int64
28	TotalWorkingYears	2149 non	null	int64
29	TrainingTimesLastYear	2149 non	null	int64

30 WorkLifeBalance	2149 non	null	int64
31 YearsAtCompany	2149 non	null	int64
32 YearsInCurrentRole	2149 non	null	int64
33 YearsSinceLastPromotion	2149 non	null	int64
34 YearsWithCurrManager	2149 non	null	int64

CLASIFICACIÓN DE VARIABLES.

Variable dependiente : Attrition (Abandono)

Variable independiente (31 variables diferentes en total) :

Información básica: Edad, sexo, educación, campo educativo, estado civil, distancia del domicilio

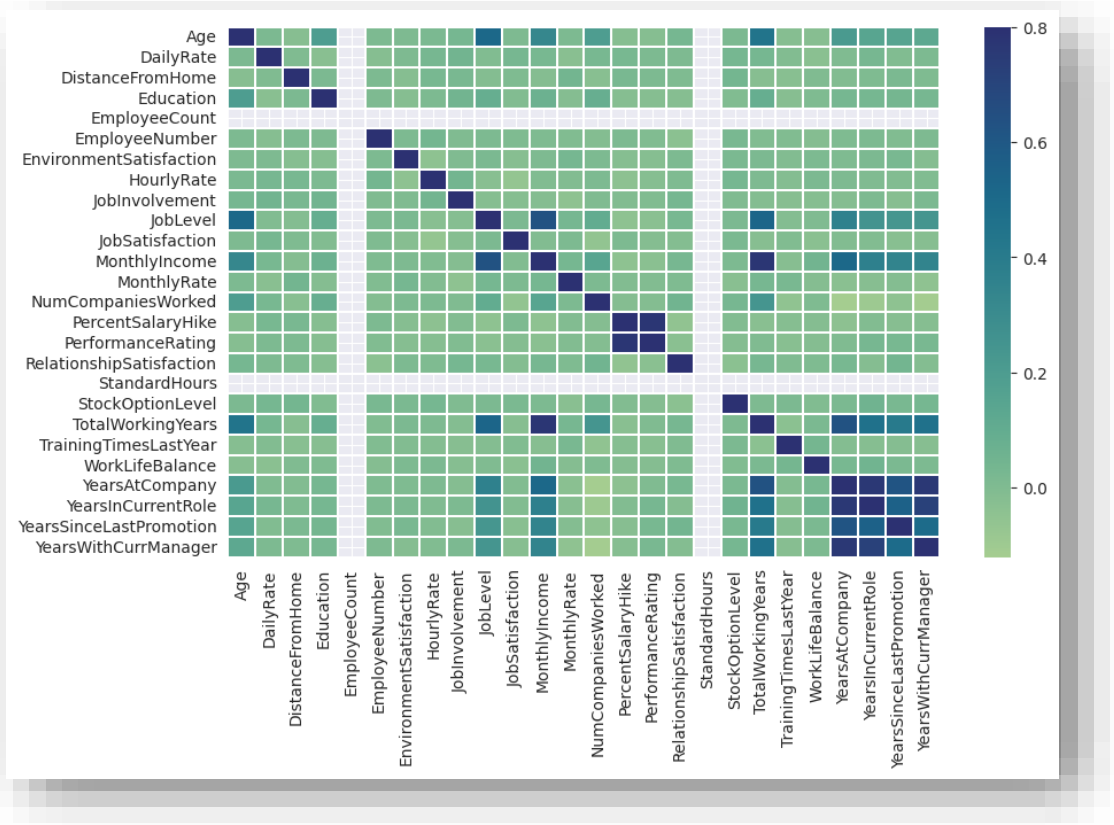
Información laboral: Departamento, Función, Nivel de puesto, Horas extras, Viajes de trabajo, Valoración del rendimiento, Nivel de opciones sobre acciones, Implicación en el trabajo

Satisfacción: Equilibrio trabajo-vida privada, Satisfacción en el trabajo, Satisfacción en las relaciones, Satisfacción con el entorno

Salario: Ingresos mensuales, tarifa mensual, tarifa diaria, tarifa por hora, porcentaje de aumento salarial

Tiempo de trabajo: Años totales de trabajo, Tiempo de formación el último año, Años en la empresa, Años en el puesto actual, Años desde el último ascenso, Años con el jefe actual, Número de empresas en las que ha trabajado.

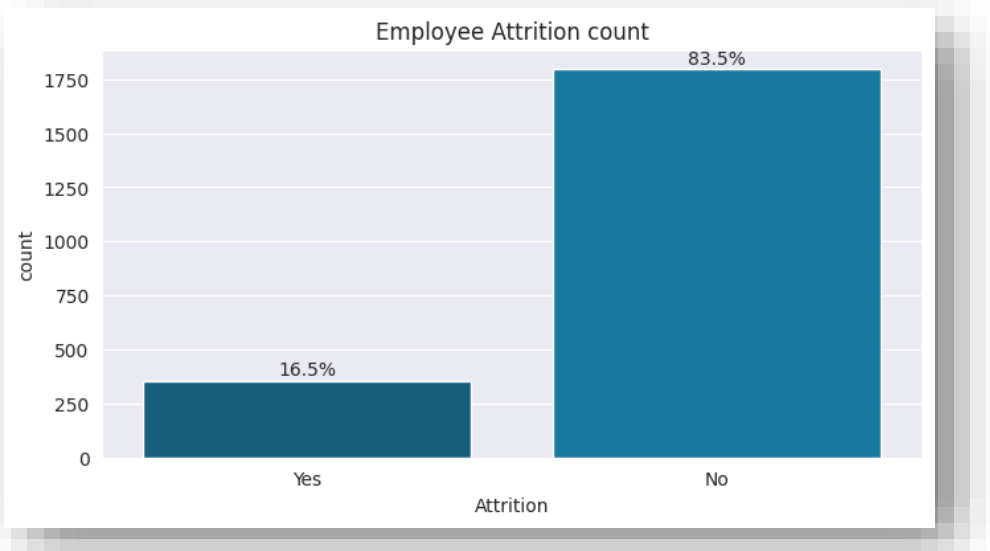
RELACIÓN DE COLUMNAS:



Los años trabajados tiene una relación positiva con el nivel de empleo y los ingresos mensuales.

También podemos ver una buena relación entre los años que tienen en la empresa, el puesto actual y los años con el directivo actual.

ANÁLISIS DE TARGET (ATTRITION).

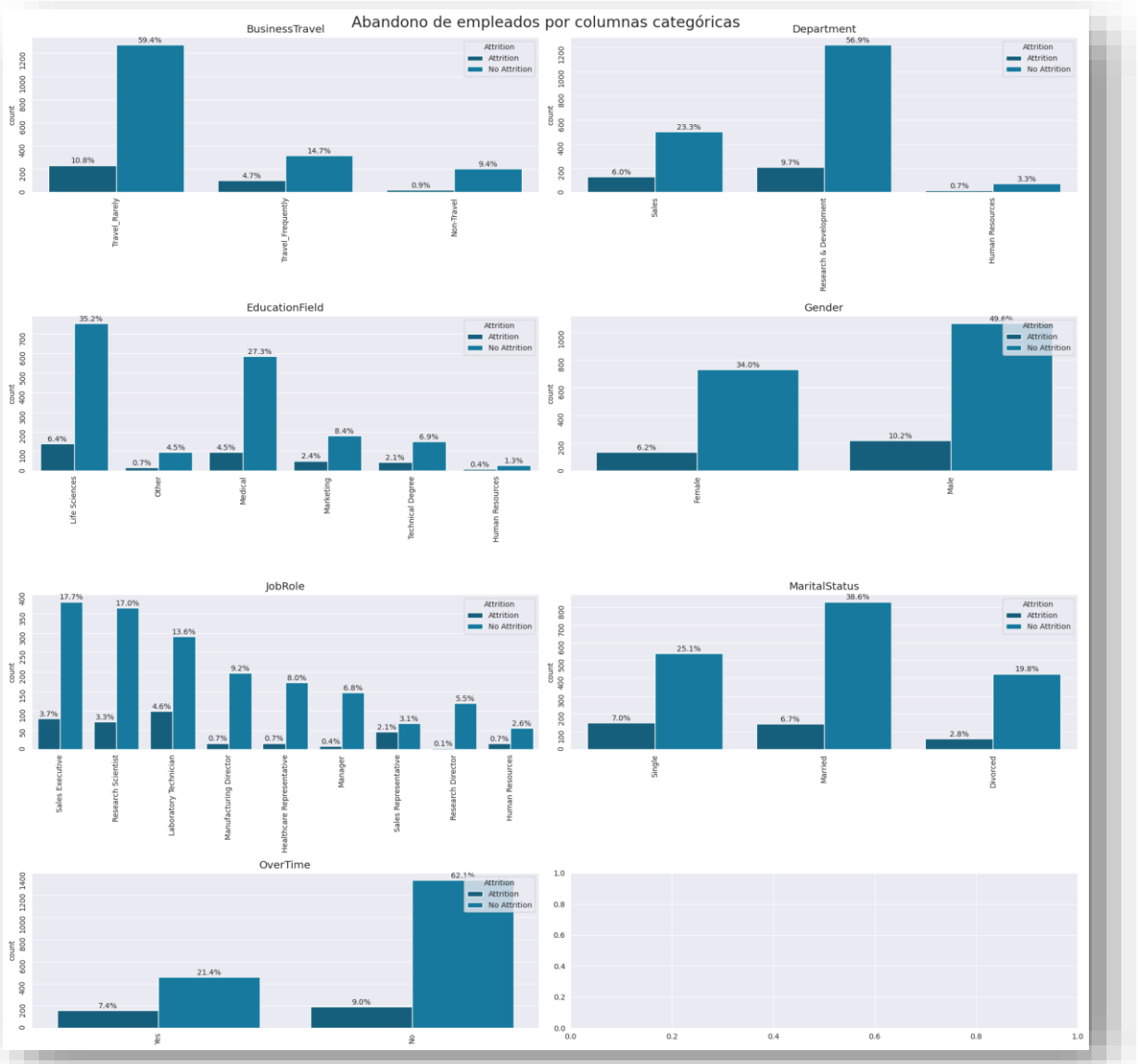


El 16,5% de los empleados han abandonado la empresa. Nuestro Target esta desequilibrada, lo que podría ser un problema para el conjunto de datos, ya que está claramente sesgado a favor de los empleados que optan por permanecer en la empresa.

DISTRIBUCIÓN DE LAS VARIABLES NUMÉRICAS.

Podemos visualizar datos que no serán útiles como “EmployeeNumber.

DISTRIBUCIÓN DE VARIABLES CATEGÓRICAS



CONVERSIÓN DE VARIABLE TARGET.

Para obtener un mejor desempeño se toma la variable Target, y se hace una conversión a variable numérica.

Conversión de variable Target

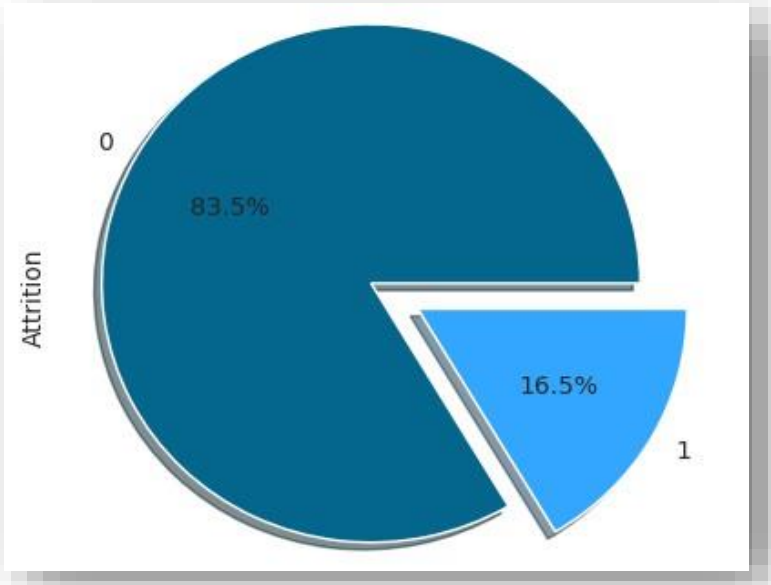
Yes = 1

No = 0

```
[ ]  df_empleados.Attrition.replace(to_replace = dict(Yes = 1, No = 0), inplace = True)
```

```
[ ]  print("Primeras 5 filas de Target")
     df_empleados[target].head()
```

```
Primeras 5 filas de Target
0    1
1    0
2    1
3    0
4    0
Name: Attrition, dtype: int64
```



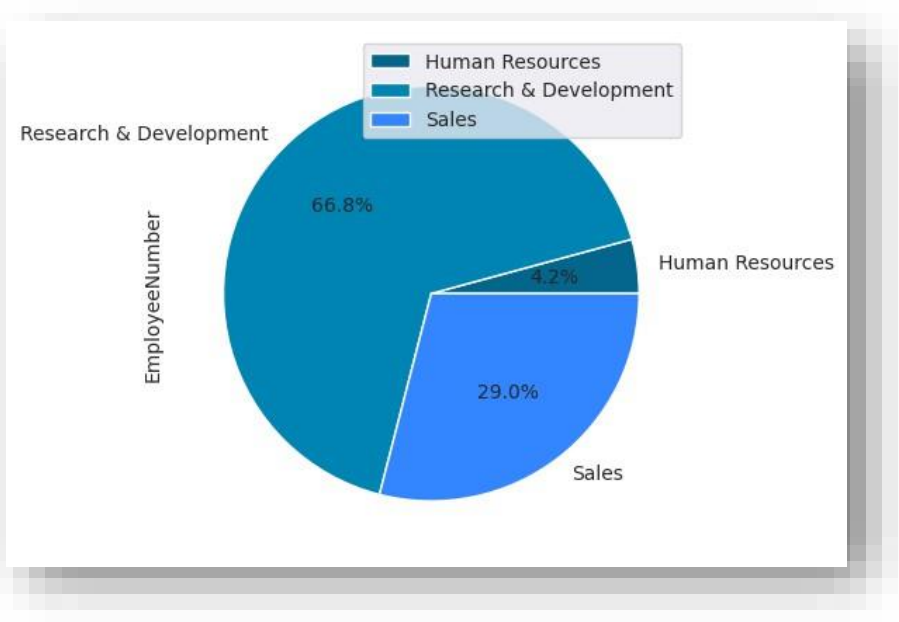
OTROS CAMBIOS REALIZADO A LOS DATOS

TRANSFORMAMOS LOS DATOS DE LAS COLUMNAS= EDUCATION, ENVIRONMENTSATISFACTION, JOBINVOLVEMENT, JOBSATISFACTION, PERFORMANCERATING, RELATIONSHIPSATISFACTION, WORKLIFEBALANCE.

```
educacion = {1:'Below College', 2:'College', 3:'Bachelor', 4:'Master', 5:'Doctor'}
df_empleados = df_empleados.replace({'Education':educacion})
```

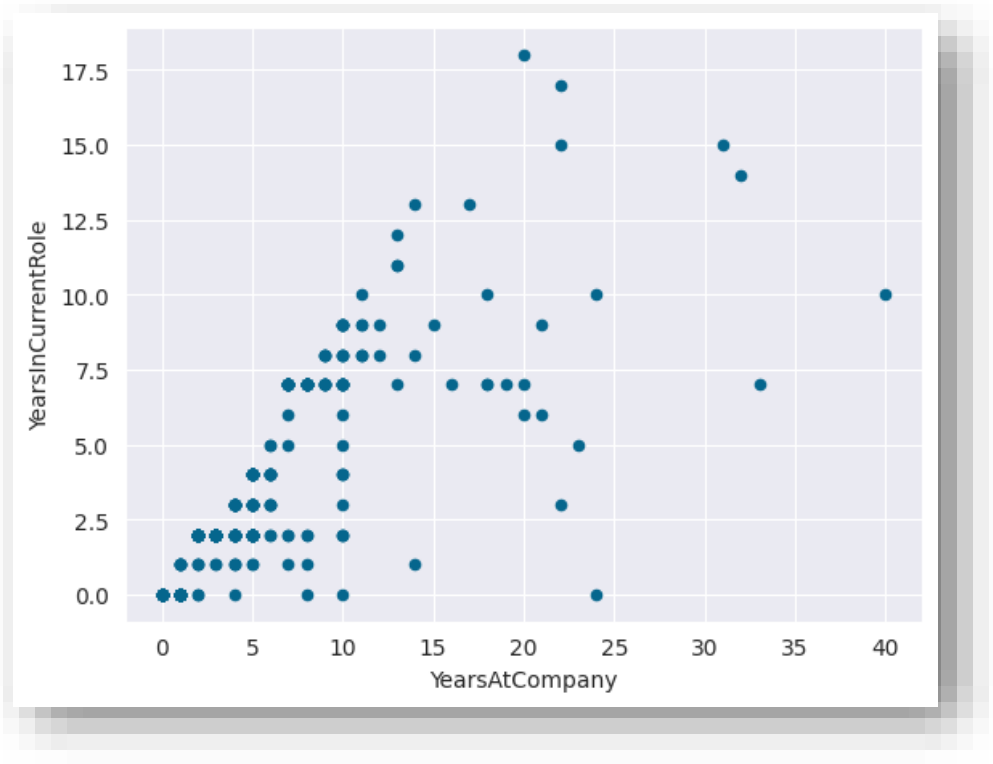
```
Satisfacción_Ambiental = {1:'Low',2:'Medium',3:'High',4:'Very High'}
df_empleados = df_empleados.replace({'EnvironmentSatisfaction':Satisfacción_Ambiental})

Implicación_en_el_trabajo = {1:'Low',2:'Medium',3:'High',4:'Very High'}
df_empleados = df_empleados.replace({'JobInvolvement':Implicación_en_el_trabajo})

Satisfacción_laboral = {1:'Low',2:'Medium',3:'High',4:'Very High'}
df_empleados = df_empleados.replace({'JobSatisfaction':Satisfacción_laboral})

Clasificación_de_Rendimiento = {1:'Low',2:'Good',3:'Excellent',4:'Outstanding'}
df_empleados = df_empleados.replace({'PerformanceRating':Clasificación_de_Rendimiento})

Satisfacción_de_la_relación = {1:'Low',2:'Medium',3:'High',4:'Very High'}
df_empleados = df_empleados.replace({'RelationshipSatisfaction':Satisfacción_de_la_relación})

Equilibrio_trabajo_vida = {1:'Bad',2:'Good',3:'Better',4:'Best'}
df_empleados = df_empleados.replace({'WorkLifeBalance':Equilibrio_trabajo_vida})

#Mostramos los cambios
df_empleados
```

VEMOS LA RELACIÓN ENTRE LOS EMPLEADOS Y LOS DEPARTAMENTOS.



Un 68% trabaja en el Departamento de Investigación y Desarrollo, un 29% en Ventas y el 4% en RRHH.

QUE OCURRE CON LA RELACIÓN YEARSATCOMPANY(AÑOS EN LA EMPRESA) Y YEARSINCURRENTROLE(AÑOS EN EL PUESTO ACTUAL).



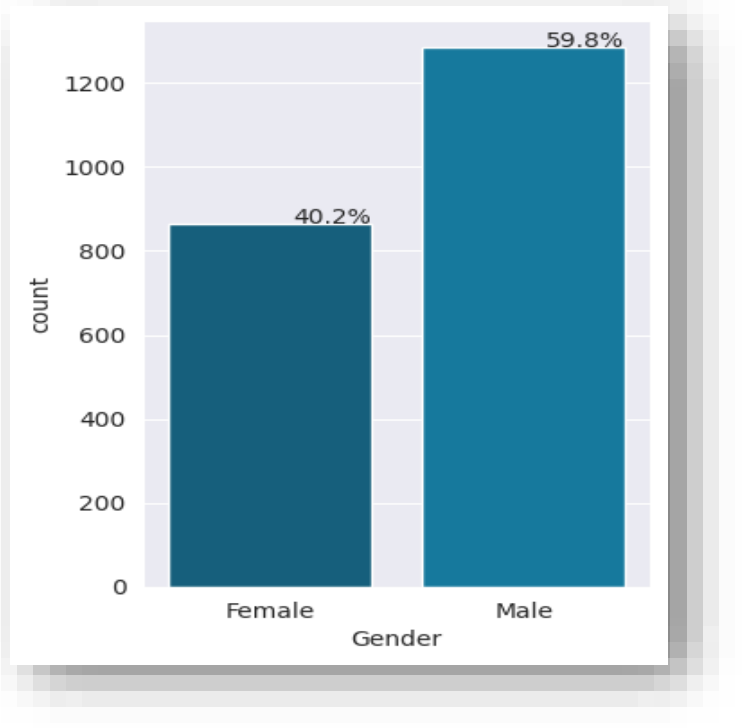
Con el gráfico de dispersión se puede ver la intensidad de la relación entre los años que el empleado pertenece a la compañía en comparación con los años que tiene la misma posición, tomando los empleados que sienten un desgaste en su trabajo. Como se puede observar vemos la mayor intensidad entre los 5 y 10 años que pertenecen a la compañía.

EDA.

Exploramos el conjunto de datos, observamos la distribución de características, cómo de correlaciones entre ellas, y creamos algunas visualizaciones.

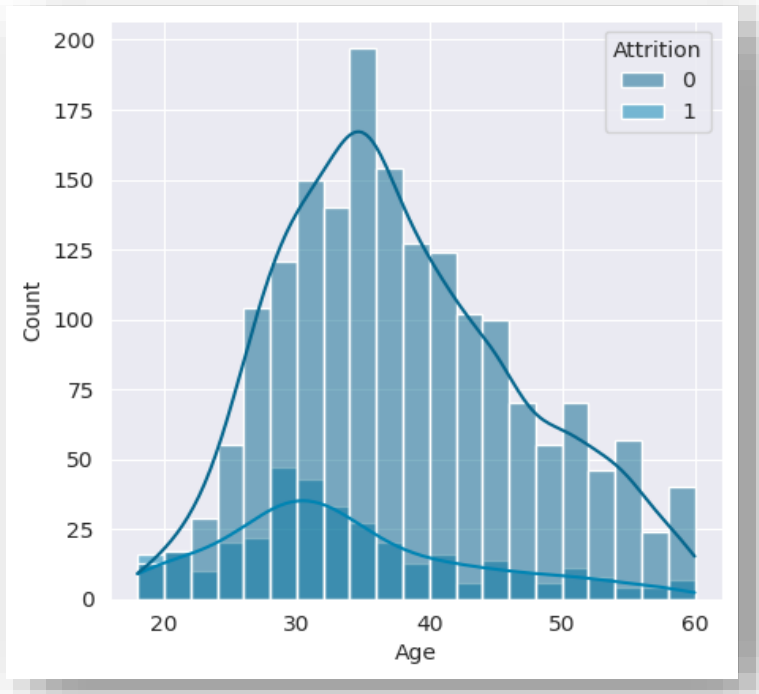
En esta instancia se hacen análisis Univariados (analizamos individualmente las variables), Bivariado (con la intención de encontrar correlaciones entre las variables), y Multivariado.

GENDER(GENERO) VS ATTRITION.



Se ve que los hombres tienen más posibilidad de Attrition, y que es mayor el número de Hombres que de mujeres.

DISTRIBUCIÓN POR LA EDAD.

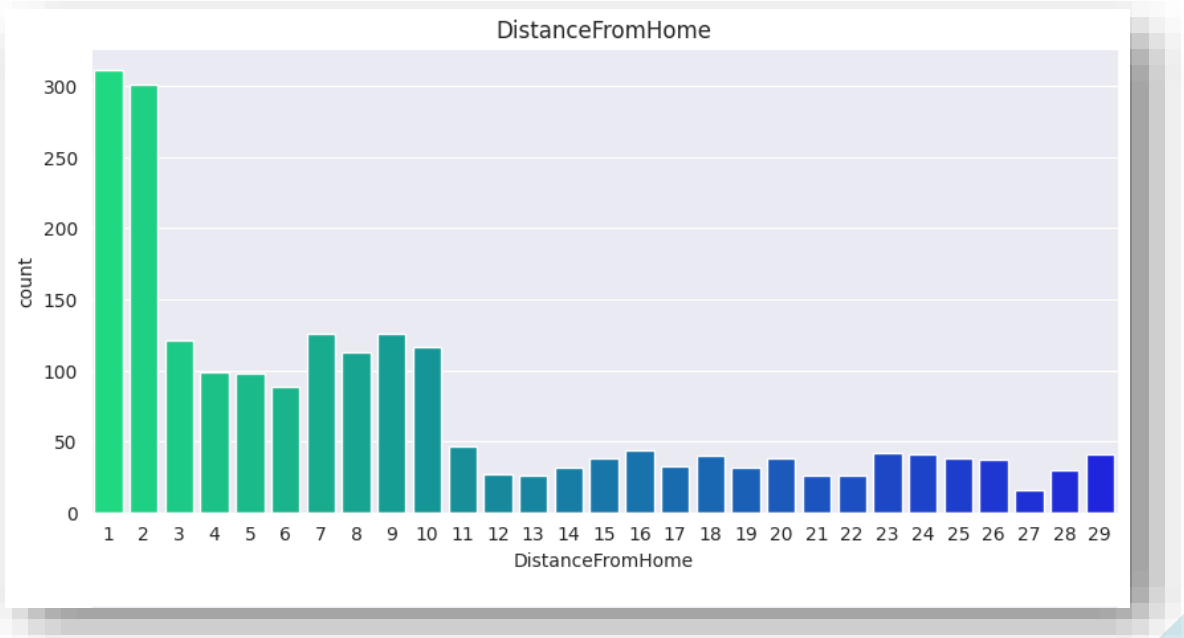


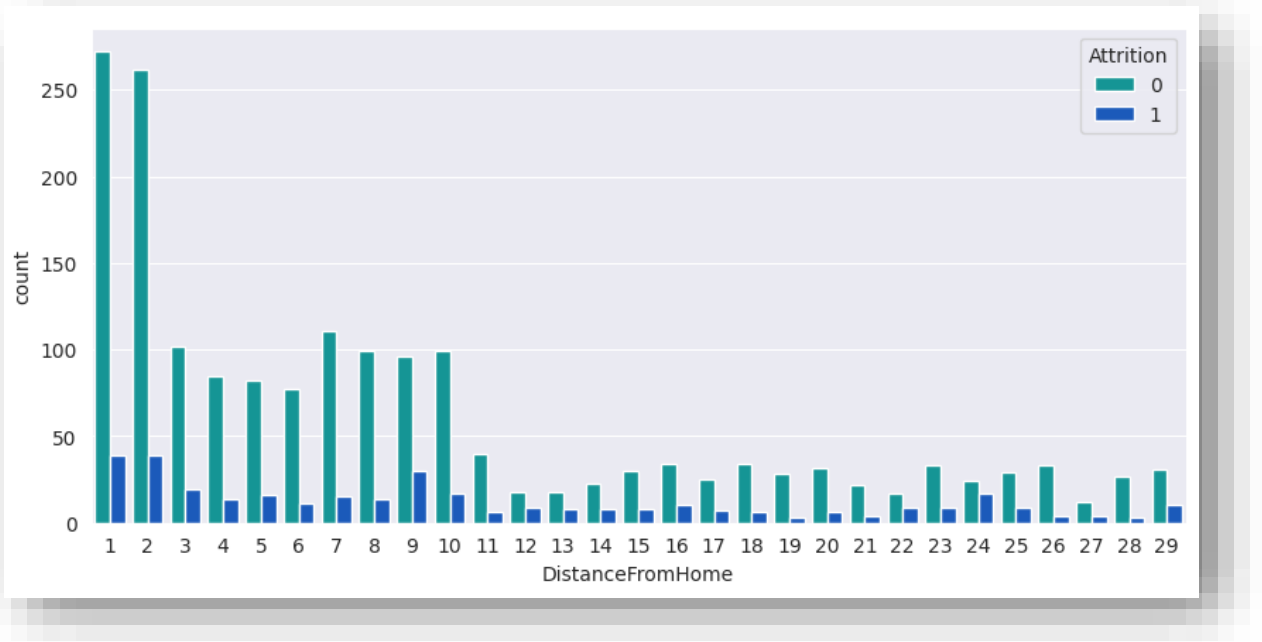
La mayoría de los empleados tienen entre 28 y 36 años, mientras que el rango de edad es de 18 a 60 años.

La empresa parece depender en gran medida del segmento de empleados menores de 40 años. Y hay un patrón cercano en el que los empleados más jóvenes tienen una mayor probabilidad de desgaste hasta alrededor de los 35 años.

LA DISTANCIA AL DOMICILIO.

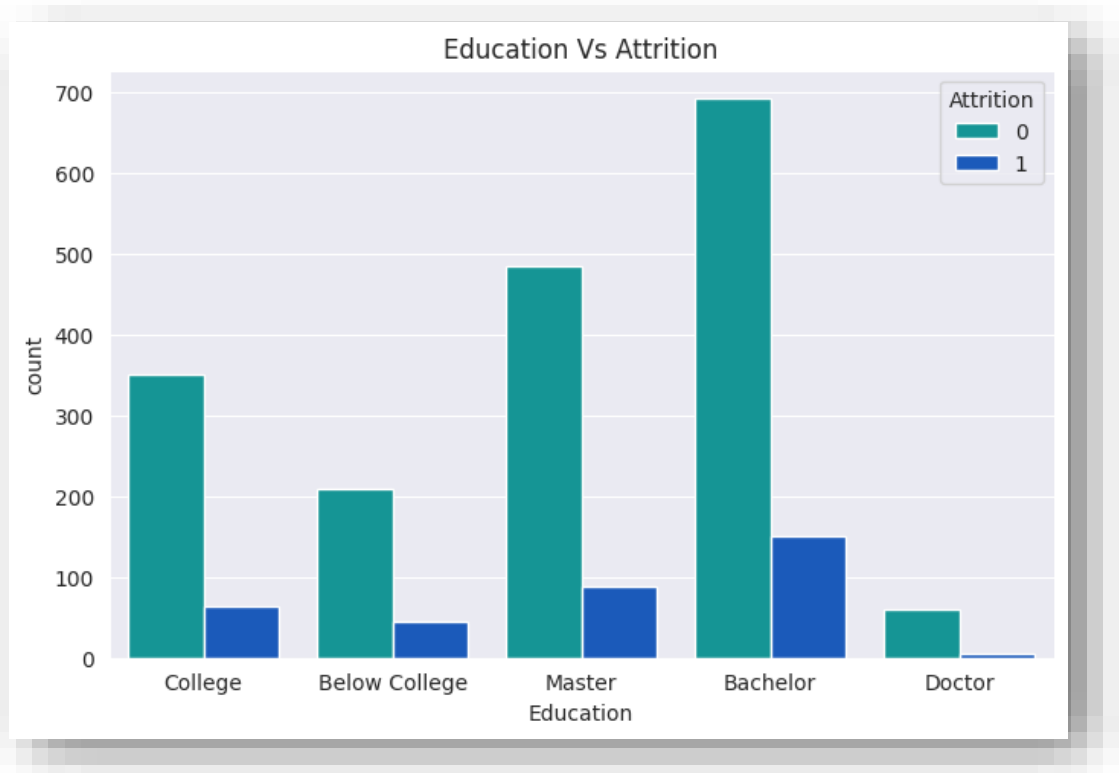
DistanceFromHome(Distancia al domicilio) vs Attrition.





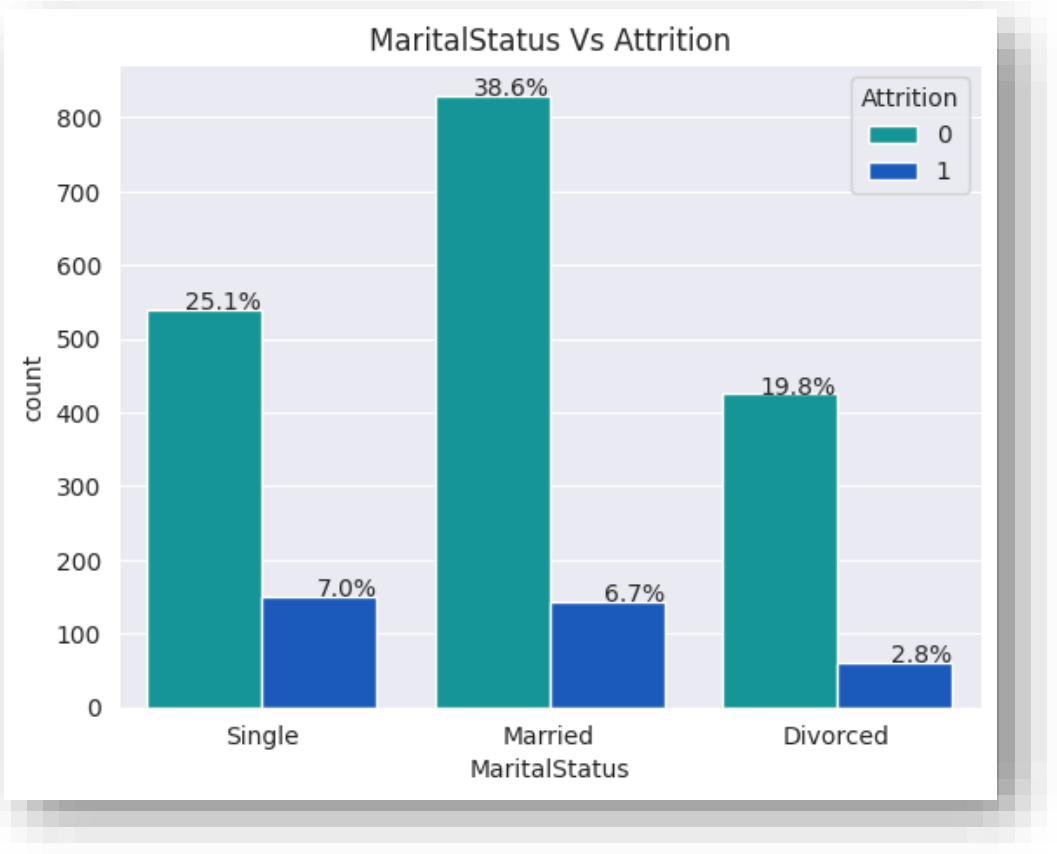
La mayoría de los empleados se encuentran a una distancia entre 1 y 10 km. Los empleados tienden a Attrition cuando la distancia a su domicilio es superior a 10 km.

EDUCATION(NIVEL DE EDUCACIÓN) VS ATTRITION.



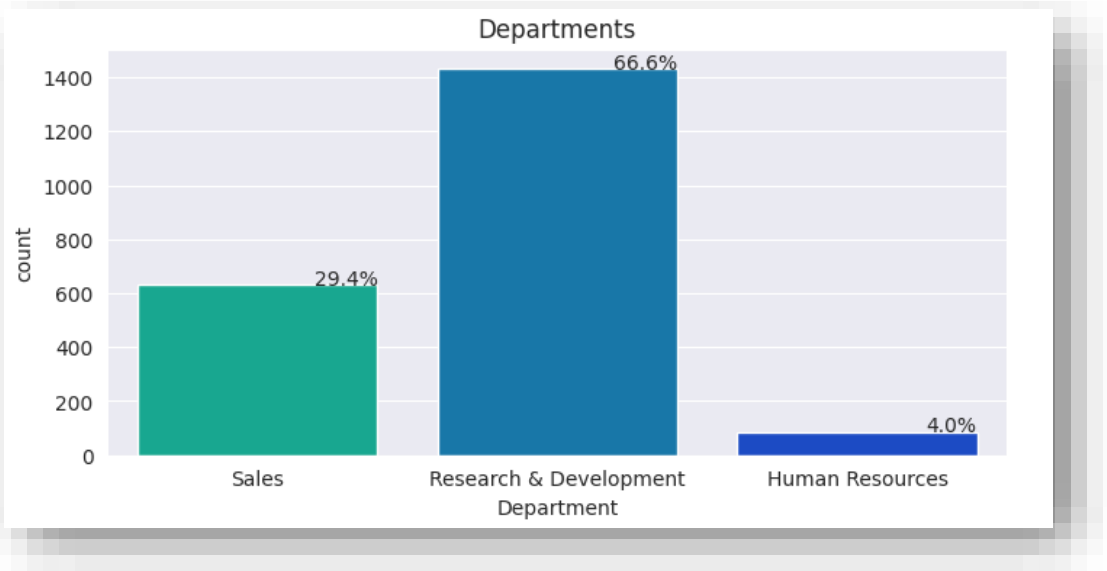
Quienes tienen el nivel de educación "Bachelor", tienen mayor probabilidad de Attrition.

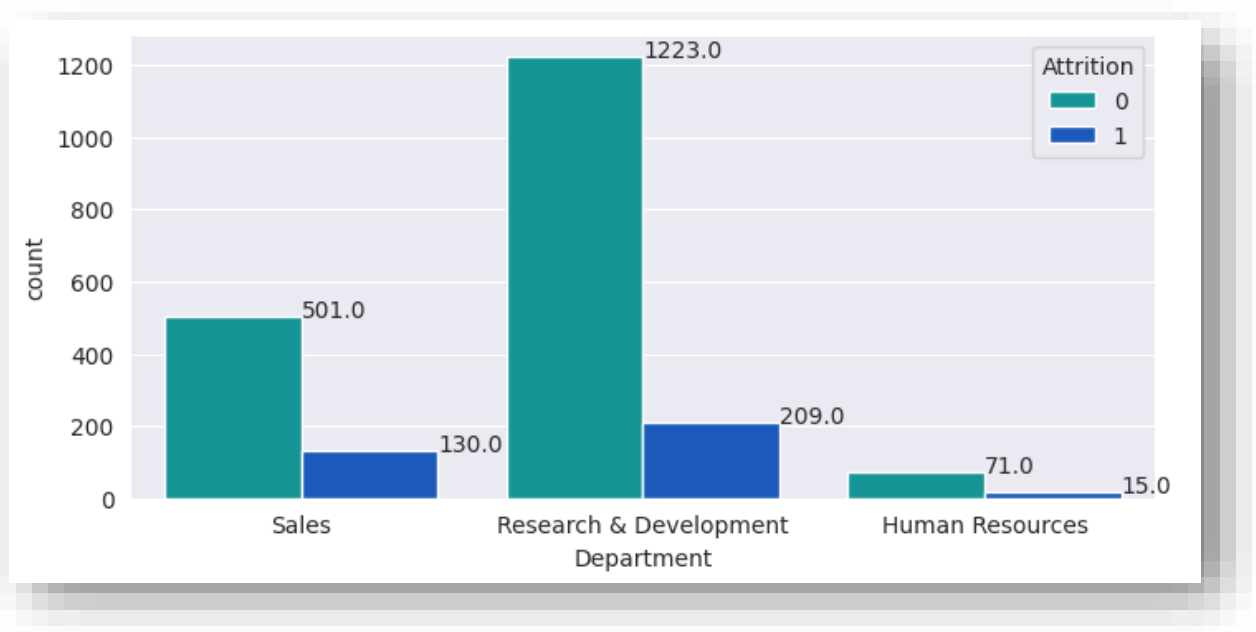
MARITALSTATUS(ESTADO CIVIL) VS ATTRITION.



El empleado cuyo estatus es "Single" (solteros) tiene claros indicios de Attrition, mientras que los demás, que están casados tienden a ser estables.

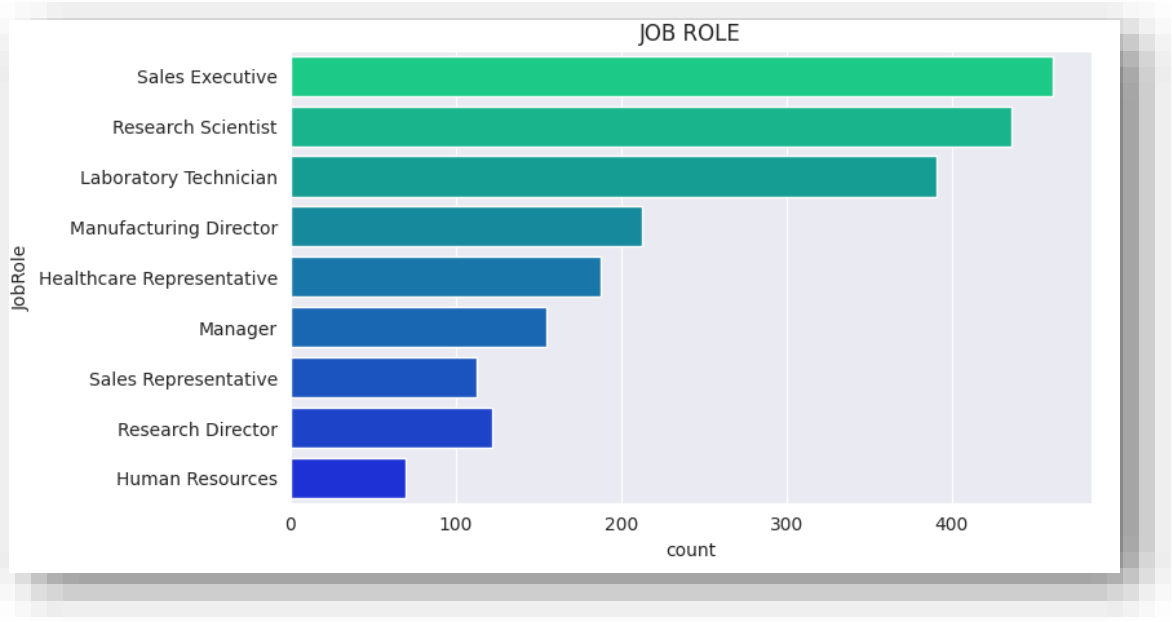
DEPARTMENT(DEPARTAMENTO) VS ATTRITION.

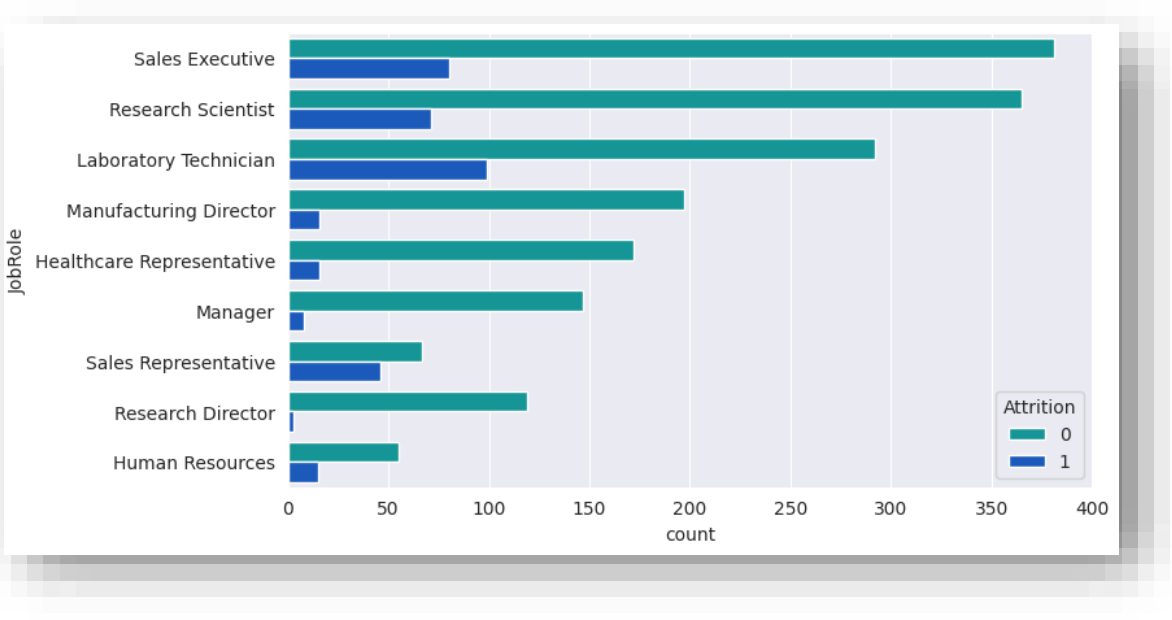




La tasa de abandono en el Dpto. de Ventas es la más alta, superando al Dpto. de R&D y RRHH.

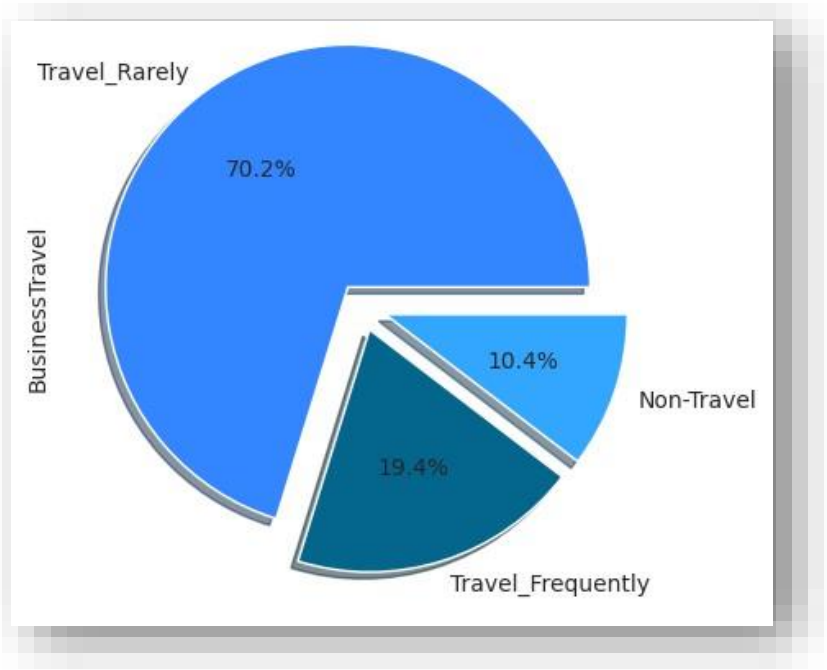
JOB ROLE (ROL) VS ATTRITION.

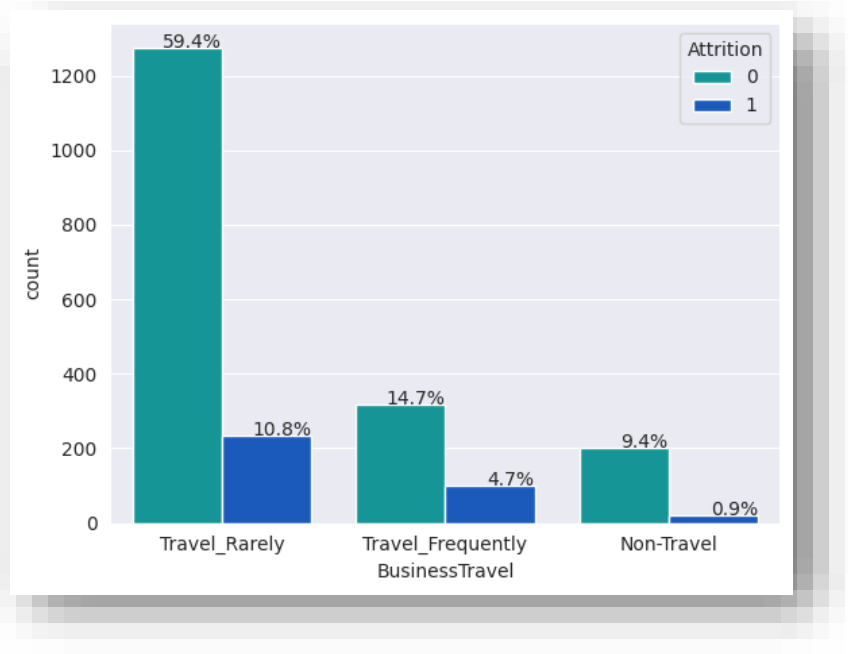




Se observar que "Sales Executive", "Sales Representative", y "Lab Technician"("ejecutivo de ventas", "representante de ventas" y "técnico de laboratorio") tienen más probabilidades de marcharse en comparación con otras funciones.

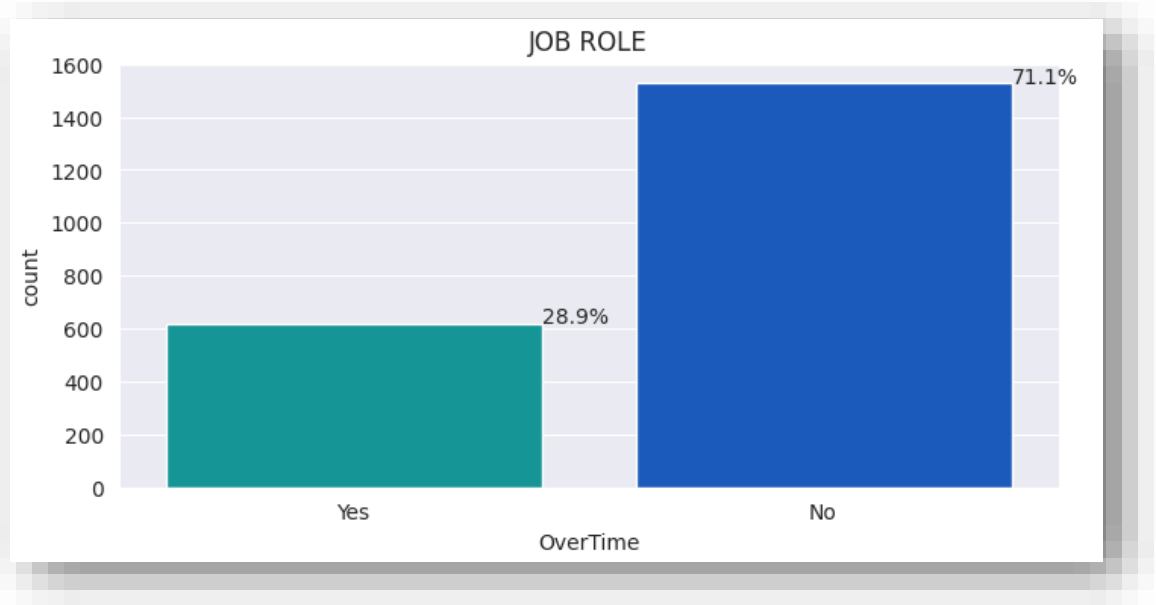
BUSINESSTRAVEL(VIAJES DE NEGOCIO) VS ATTRITION

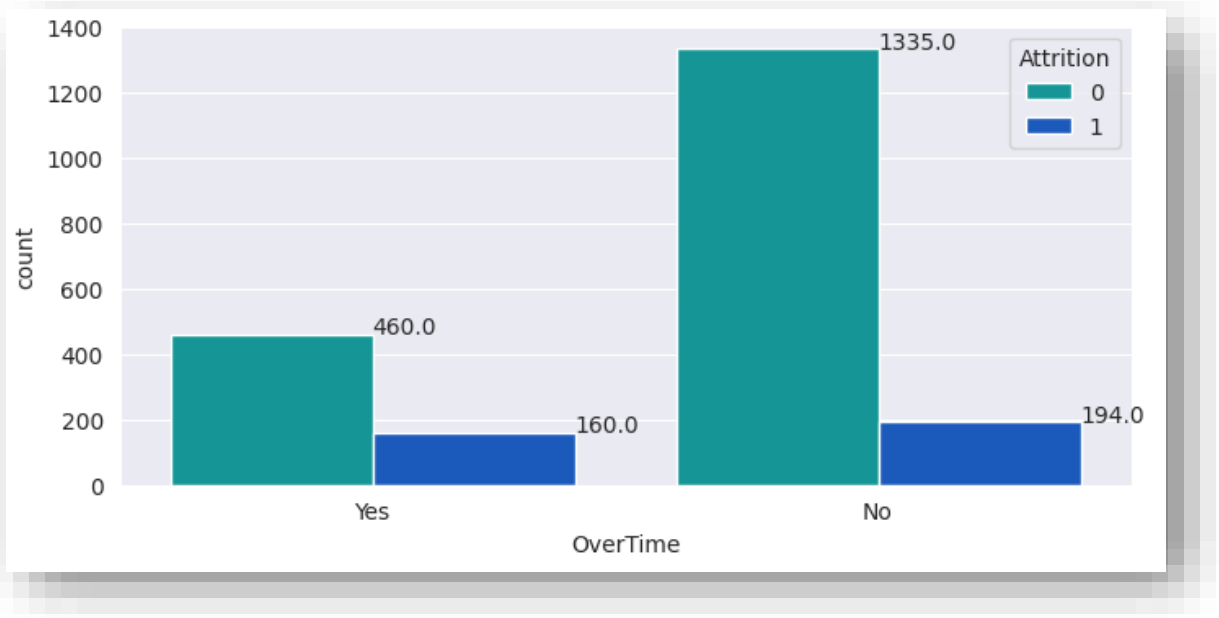




Los empleados que viajan poco son los que presentan un número mayor de Attrition, mientras que los empleados que no tienen que viajar son los que presentan menos posibilidades de Attrition.

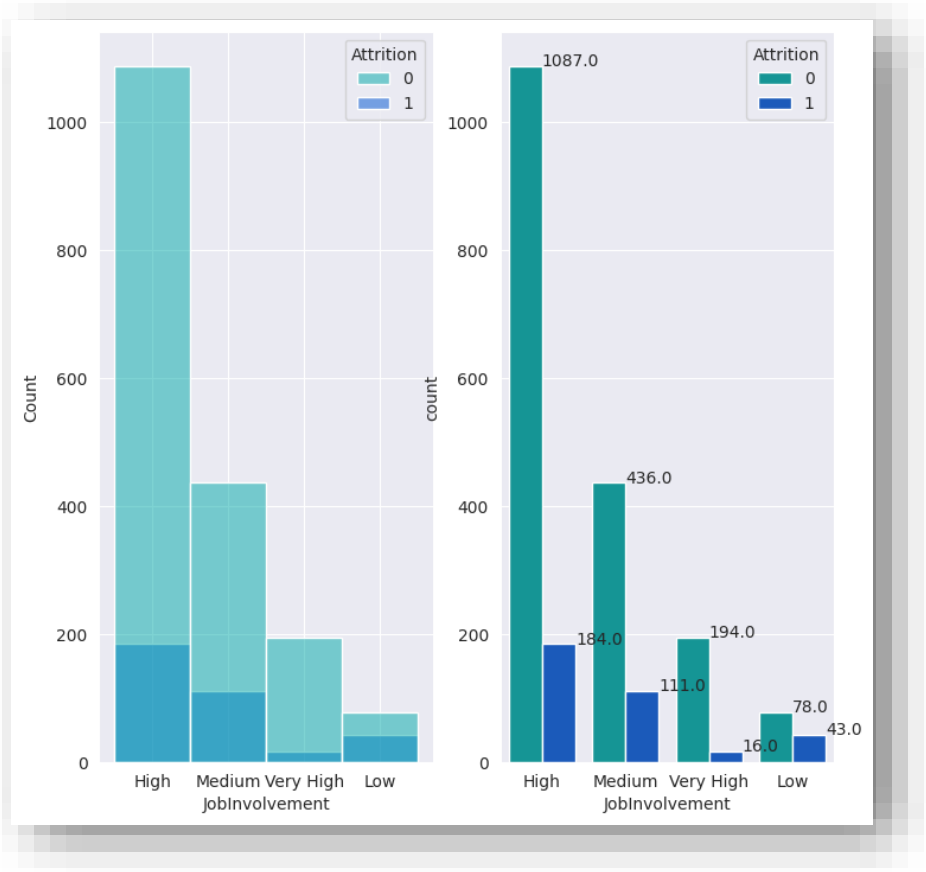
OVERTIME(HORAS EXTRAS) VS ATTRITION.





A los empleados que permanecen en la empresa no se les exigió que hicieran horas extras, mientras que a los empleados que se han ido se les pidió que hicieran. Se puede ver y considerar una causa de Attrition.

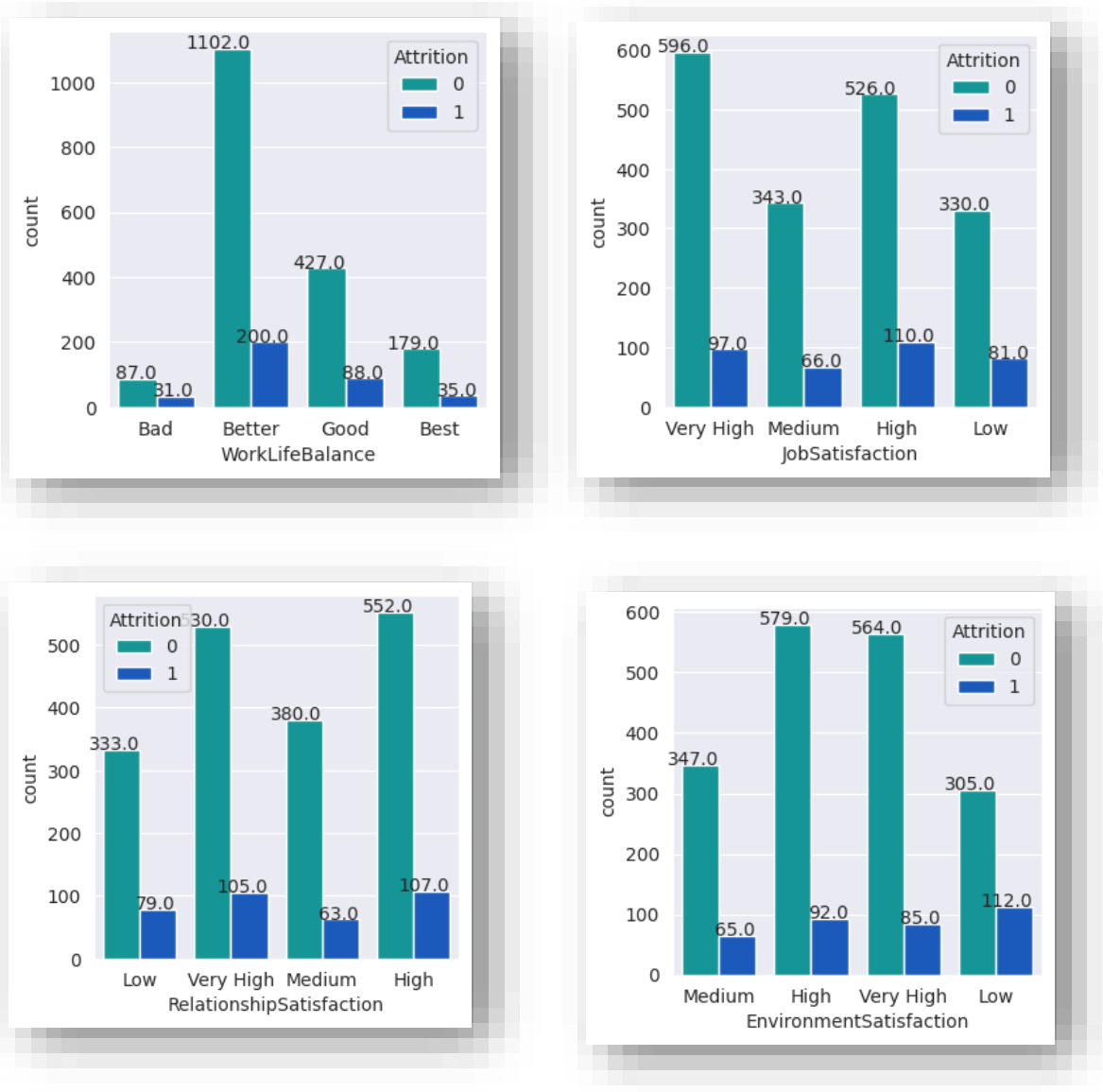
JOBINVOLVEMENT(PARTICIPACIÓN) VS ATTRITION.



La mayoría de los empleados consideran que su implicación en el trabajo es High. La tasa de abandono de los empleados que consideran que su implicación en el trabajo es Low(35%) es superior a High(14%).

Attrition= JobInvolvement-High(14%) / JobInvolvement-Medium(20%) / JobInvolvement-Very High(7%) / JobInvolvement-Low(35%).

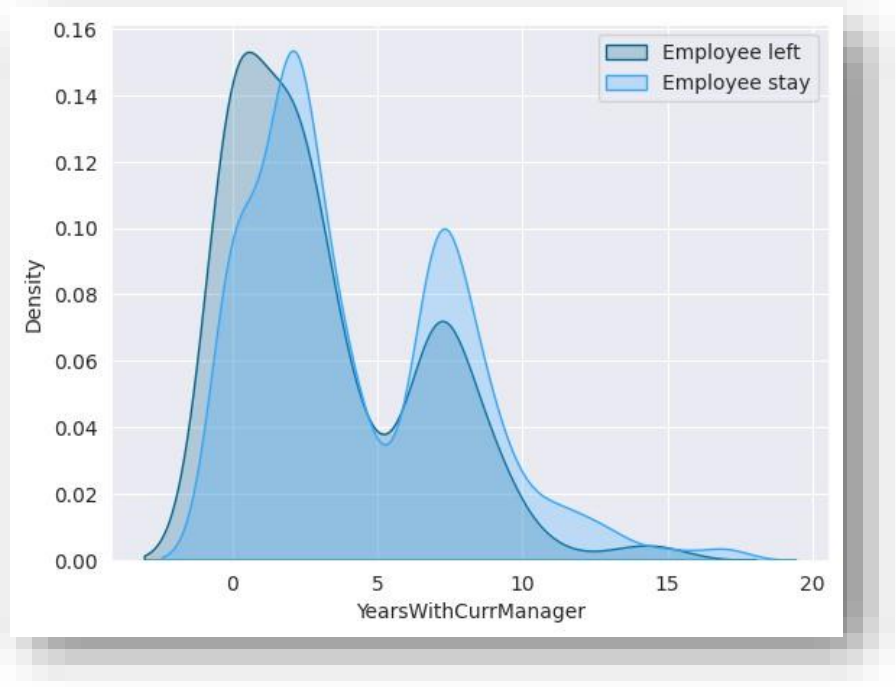
WORK-LIFE BALANCE(EQUILIBRIO VIDA-TRABAJO), JOB SATISFACTION(SASTIFACCIÓN LABORAL), RELATIONSHIP SATISFACTION(RELACIÓN SASTIFACCIÓN), ENVIROMENT SATISFACTION(SASTIFACCIÓN DEL AMBIENTE LABOTRAL) VS ATTRITION.



Los empleados que no Attrition, la satisfacción con el trabajo, las relaciones y el entorno se situó en gran medida por encima High. Los que Atttrition el equilibrio entre la vida laboral y personal se sitúa en Low.

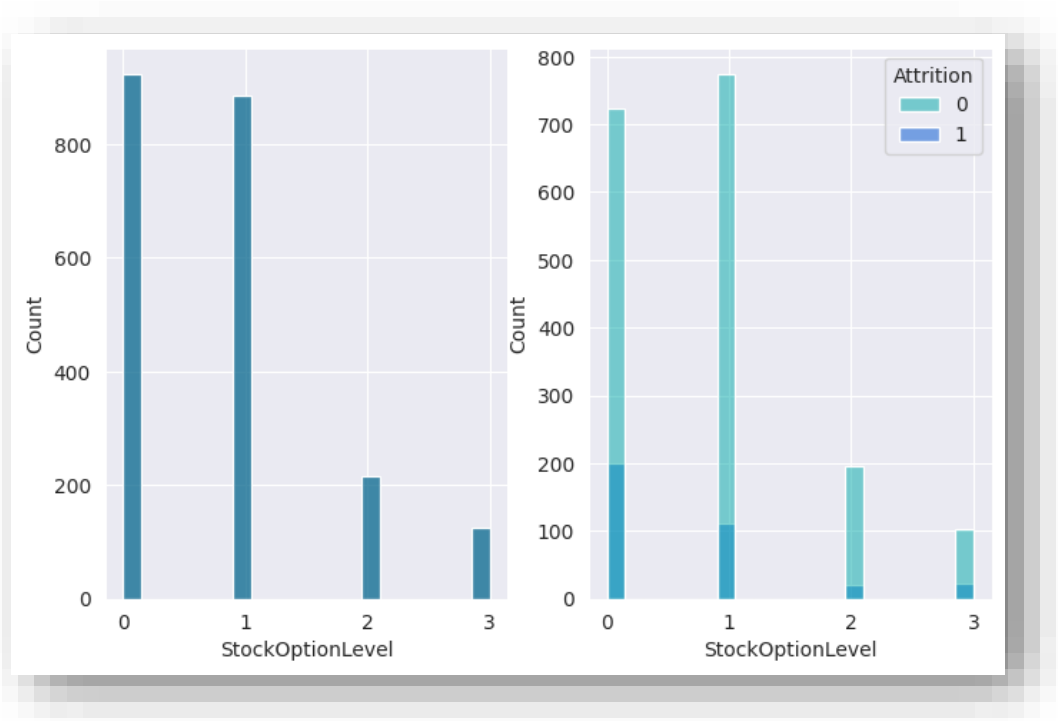
YEARS WITH CURR MANAGER(AÑOS COMO GERENTE) VS ATTRITION.

- Employee left= Empleado que se va.
- Employee stay= Emplaedo que se queda.



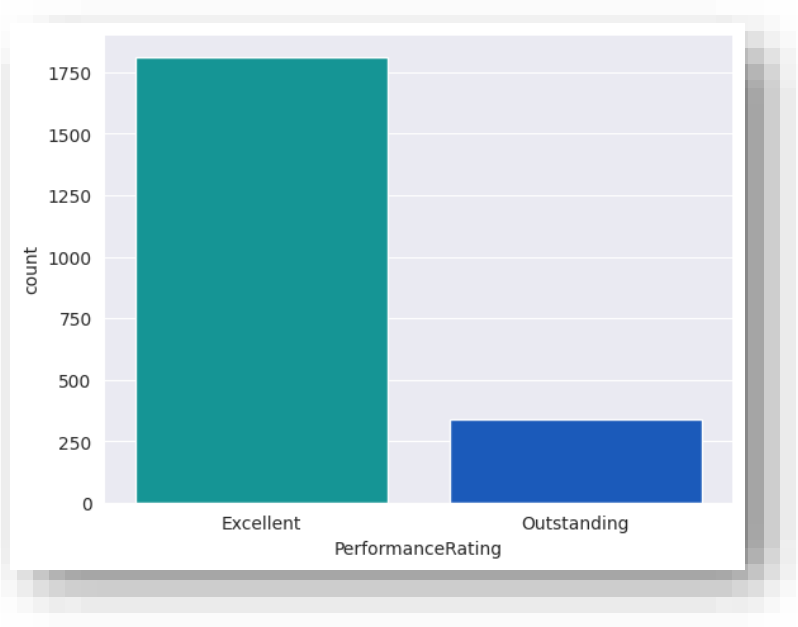
Podemos observar que los empleados que tienden a irse tienen menos de 2 años con el gerente actual, y tienden a quedarse si tienen más de 5 años.

STOCK OPTION LEVEL(NIVEL DE ACCIONES) VS ATTRITION.



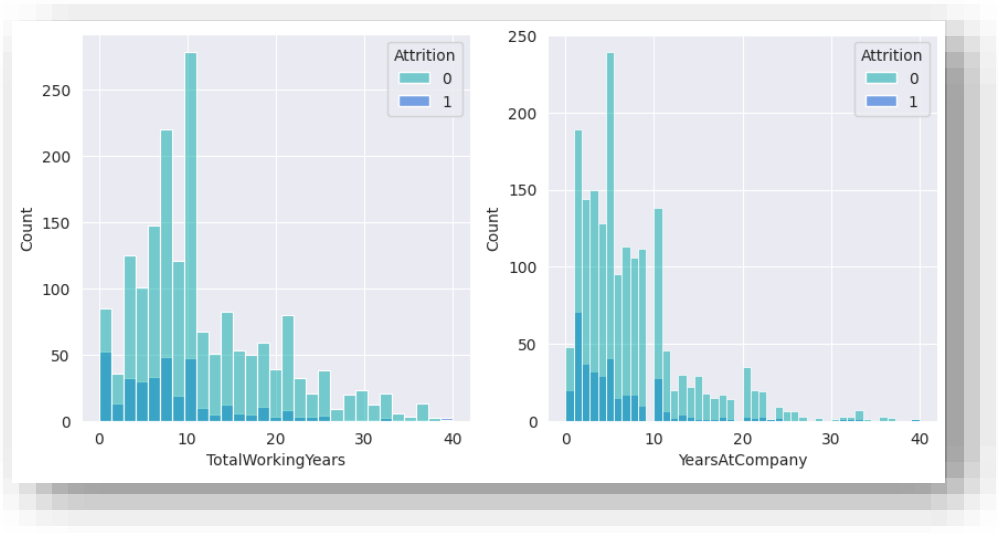
En los niveles 0 y 1 hay mayor Attrition.

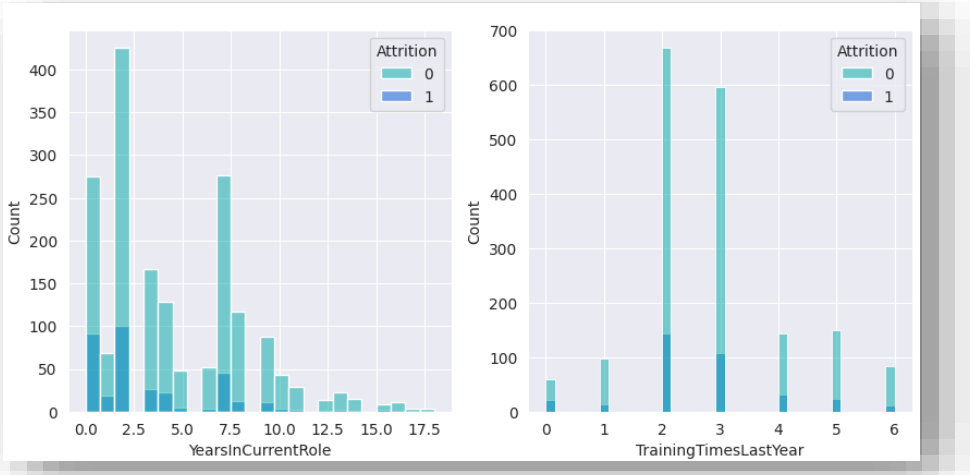
PERFORMANCE RATING (CLASIFICACIÓN DE RENDIMIENTO) VS ATTRITION.



La valoración del rendimiento de los empleados ha sido muy buena. No es determinante de Attrition.

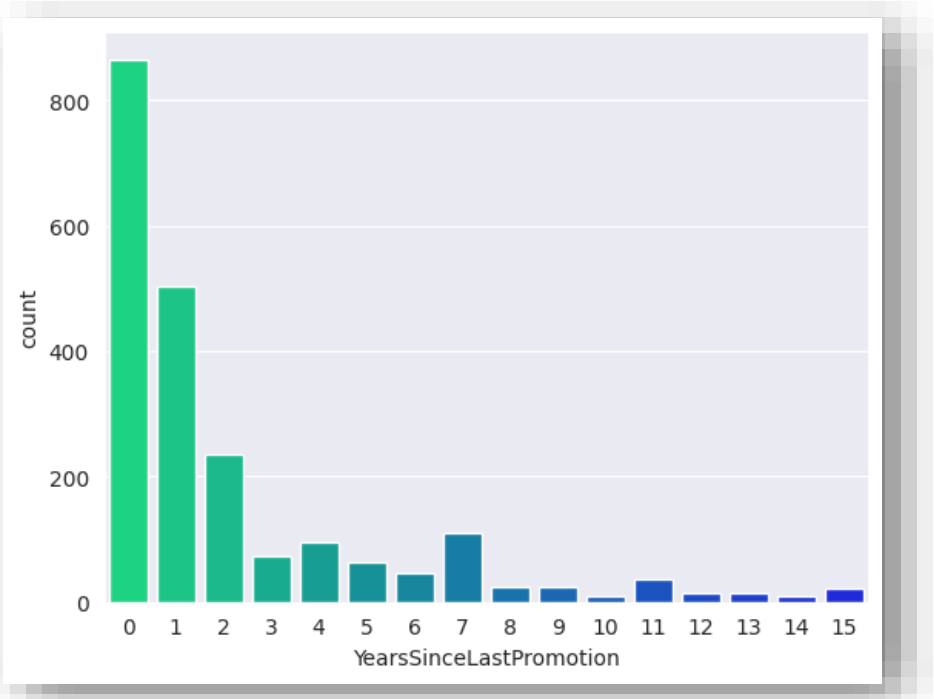
TOTALWORKINGYEARS (TOTAL AÑOS TRABAJADOS), YEARSINCURRENTROLE(AÑOS EN EL ROL ACTUAL), YEARSATCOMPANY(AÑOS EN LA EMPRESA), TRAININGTIMESLASTYEAR(TIEMPO DE ENTRENAMIENTO DEL AÑO PASADO) VS ATTTRITION.





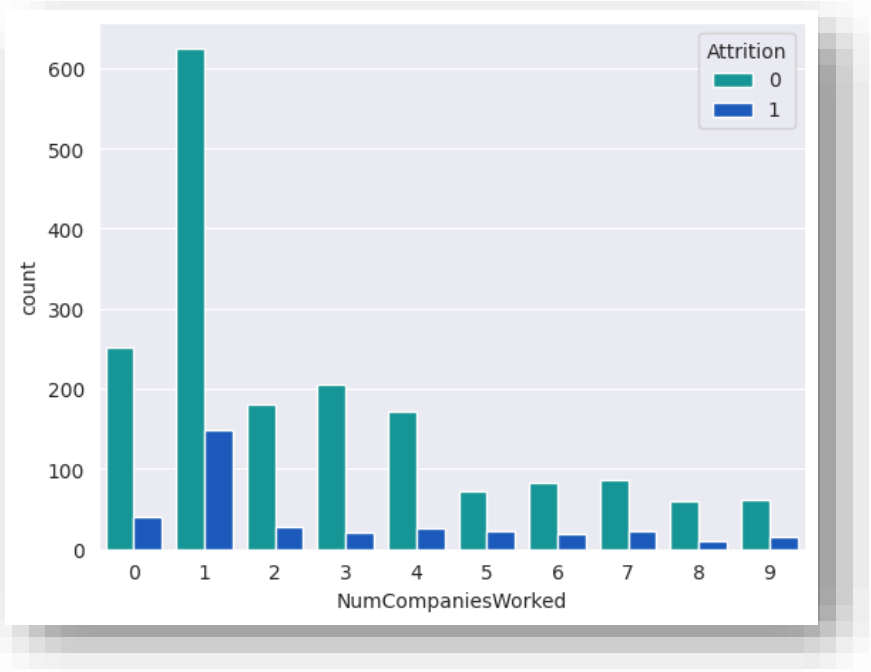
Se puede ver que más empleados tienden a irse al tener menos de 7 años de experiencia, menos de 5 años de antigüedad, y menos de 2 años en su puesto actual.

YEARSSINCELASTPROMOTION (AÑOS DESDE LA ÚLTIMA PROMOCIÓN) VS ATTRITION.



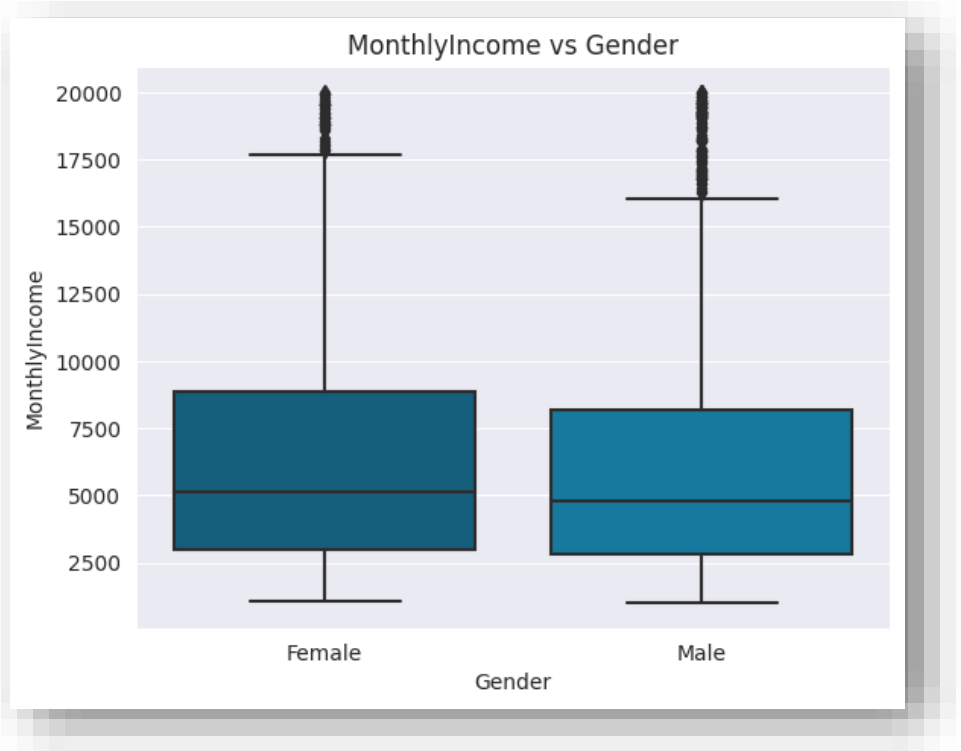
En la mayoría de los empleados su promoción ha sido hace 2años, y hay empleados con 15 en la empresa sin ninguna promoción, podemos decir que puede ser potencial Attrition.

NUMCOMPANIESWORKED (NRO EMPRESAS TRABAJADAS)



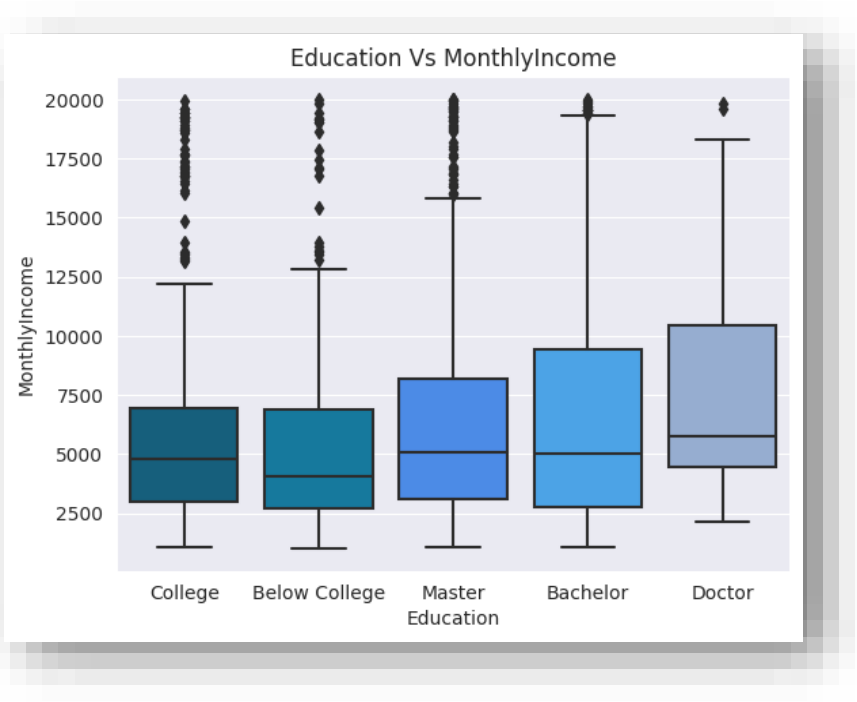
Los empleados que tienen 1 año de experiencia tienen mayor Attrition.

MONTHLYINCOME (INGRESOS MENSUALES) VS GENDER (GENERO).



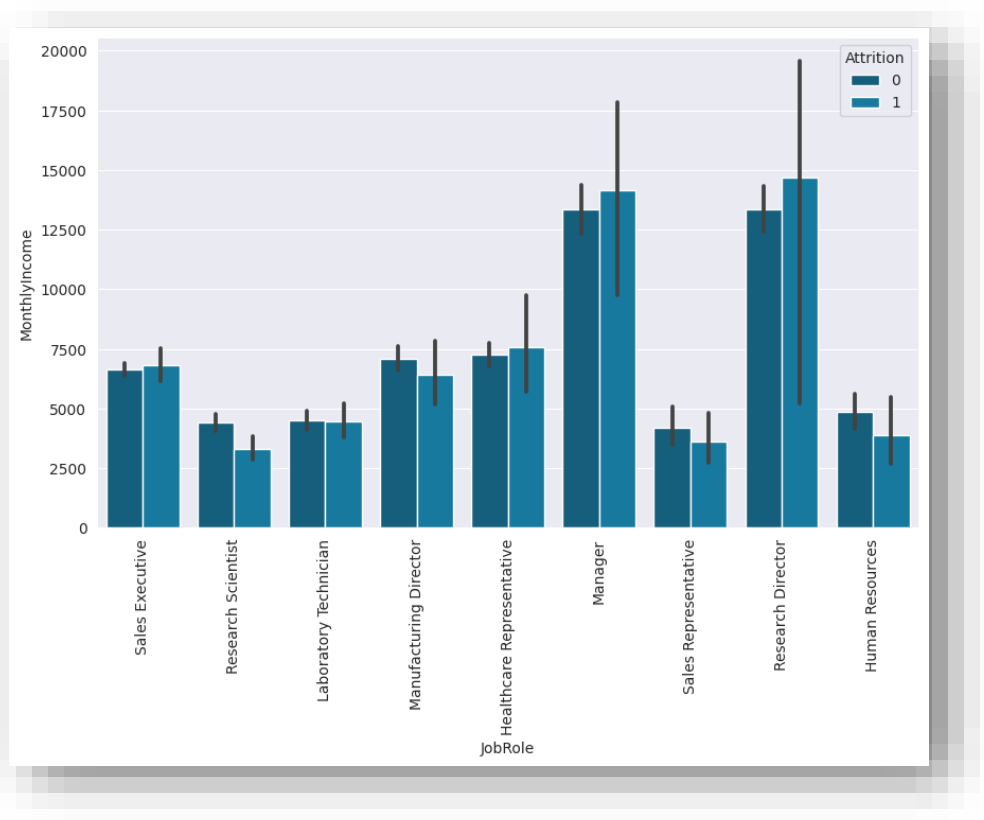
Podemos ver que las mujeres y los hombres muestran una variabilidad similar, tanto en la media, la mediana y el cuartil.

EDUCATION (EDUCACIÓN) VS MONTHLYINCOME (INGRESOS MENSUALES).



Con el nivel de estudios, aumentan los ingresos medios mensuales.

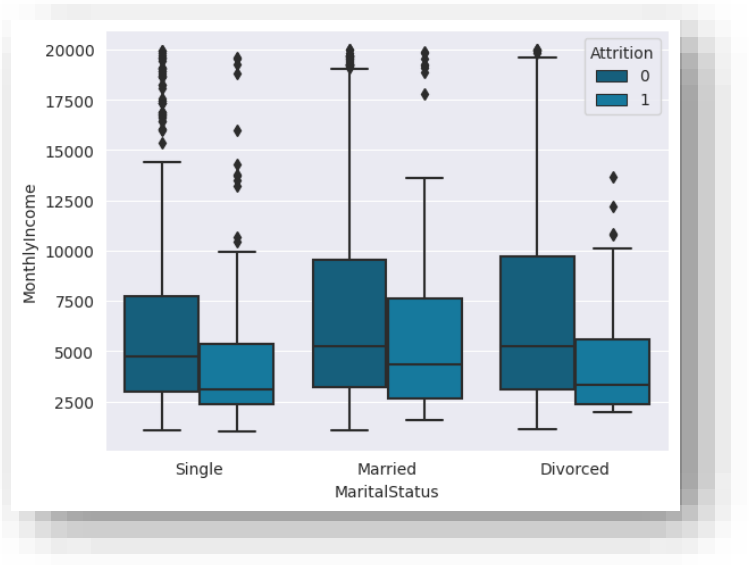
MONTHLYINCOME (INGRESOS MENSUALES) VS JOBROLE (ROL).



Podemos observar que los técnicos de laboratorio, los científicos de investigación y los representantes y ejecutivos de ventas tienen salarios muy bajos, lo que podría generar descontento, y

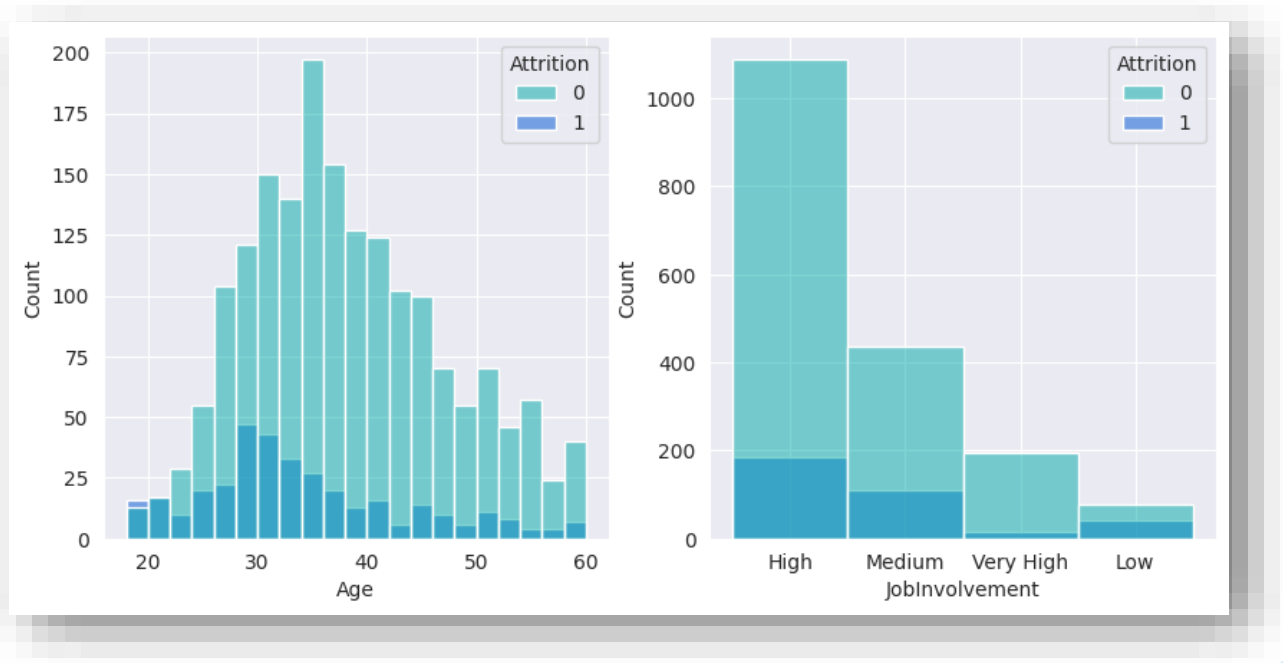
llevar a "Attrition". Y también se ve que el departamento de RRHH es el que más abandonos tiene, y también tiene salarios muy bajos.

MARITALSTATUS (ESTADO CIVIL) VS MONTHLYINCOME (INGRESOS MENSUALES).



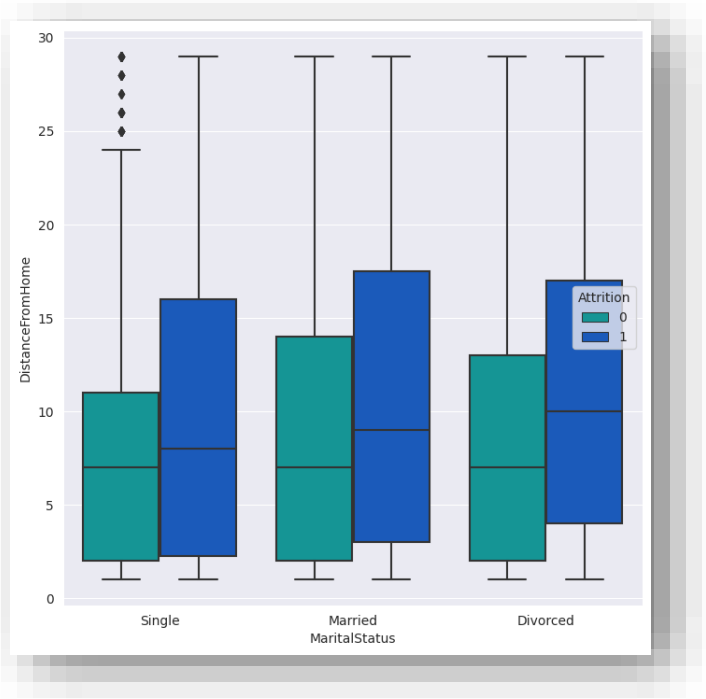
Los empleados solteros tienen ingresos más bajos, esa es la razón por la que hay gran rotación en ese rango etario.

AGE (EDAD) VS JOBINVOLVEMENT (PARTICIPACIÓN).



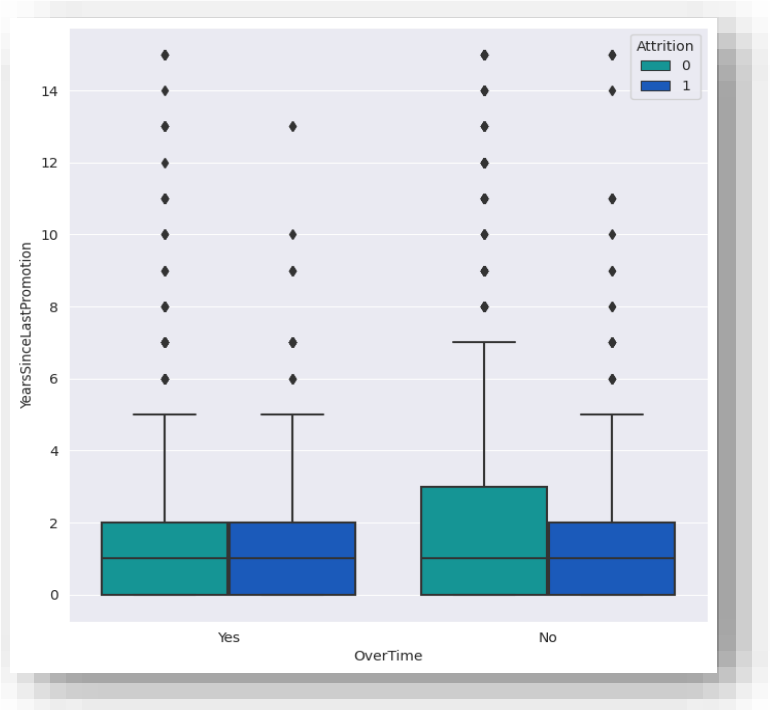
Los empleados con baja implicación en el trabajo tienden a ser más jóvenes.

MARITALSTATUS (ESTADO CIVIL) VS DISTANCEFROMHOME (DISTANCIA DESDE CASA).



La distancia desde el domicilio podría correlacionarse positivamente con un mayor nivel de Attrition. Además vemos que la mediana de los empleados que abandonan la empresa tuvo que desplazarse más que los que se quedaron, además de su estado civil.

OVERTIME (HORAS EXTRAS) VS YEARSSINCELASTPROMOTION (AÑOS DESDE LA ÚLTIMA PROMOCIÓN).



Los empleados que hacen más horas extras tardan de media más años en ser ascendidos. Esto podría desmotivar y ser razón de Attrition.

VALORES MEDIOS.

Valores medios, de todas las características para los casos de empleados que abandonan y empleados que permanecen en la empresa. Si tenemos en cuenta la edad, los valores medios de los empleados que se quedan son 37,71, es decir, más que los empleados que se van, 33,72. Del mismo modo, DailyRate y JobLevel son más altos para los empleados que se quedan que para los que se van. Los empleados que se quedan tienen valores más altos para las características : TotalWorkingYears, YearsAtCompany, YearsInCurrentRole & YearsWithCurrManager.

FEATURE ENGINEERING.

SEGUIMOS LA TRANSFORMACIÓN A NUMÉRICO, PARA UNA MAYOR EFICIENCIA.

Utilizamos LabelEncoder, asignamos una etiqueta numérica única a cada categoría o clase única en la variable.

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
for feat in obj_dtypes:
    df_empleados[feat] = le.fit_transform(df_empleados[feat].astype(str))
print (df_empleados.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2149 entries, 0 to 2148
Data columns (total 32 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Age                      2149 non-null   int64
 1   Attrition                2149 non-null   int64
 2   BusinessTravel           2149 non-null   int64
 3   DailyRate                2149 non-null   int64
 4   Department               2149 non-null   int64
 5   DistanceFromHome         2149 non-null   int64
 6   Education                2149 non-null   int64
 7   EducationField           2149 non-null   int64
 8   EmployeeNumber           2149 non-null   int64
 9   EnvironmentSatisfaction  2149 non-null   int64
 10  Gender                   2149 non-null   int64
```

CONSTRUCCIÓN DE MODELOS.

COMENZAMOS DIVIDIENDO LOS DATOS.

```
[ ] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state = 0, shuffle = True)

Counter(y_train)

Counter({0: 1337, 1: 274})
```

LOGISTIC REGRESSION.

```
[ ] log_reg_model = LogisticRegression(max_iter=1000, solver = "newton-cg")
    log_reg_model.fit(X_train, y_train)
```

```
LogisticRegression
LogisticRegression(max_iter=1000, solver='newton-cg')
```

```
[ ] y_pred = log_reg_model.predict(X_test)
    print("Model accruracy score: {}".format(accuracy_score(y_test, y_pred)))
```

```
Model accruracy score: 0.8698884758364313
```

	precision	recall	f1-score	support
0	0.89	0.97	0.93	458
1	0.62	0.31	0.42	80
accuracy			0.87	538
macro avg	0.76	0.64	0.67	538
weighted avg	0.85	0.87	0.85	538

Podemos ver que el modelo predice bastante bien a quienes no se van (93% de precisión), pero no predice tan bien a quienes se podrían ir (42% de precisión).

RANDOM FOREST.

```
random_forest_model = RandomForestClassifier(random_state = 0)
random_forest_model.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(random_state=0)
```

```
y_pred = random_forest_model.predict(X_test)
print("Model accruracy score: {}".format(accuracy_score(y_test, y_pred)))
```

```
Model accruracy score: 0.8661710037174721
```

	precision	recall	f1-score	support
0	0.87	0.99	0.93	458
1	0.72	0.16	0.27	80
accuracy			0.87	538
macro avg	0.80	0.58	0.60	538
weighted avg	0.85	0.87	0.83	538

Nuevamente vemos que el modelo predice bien a quienes no se van (93% de precisión), pero tiene una mala predicción de los que podrían irse (27% de precisión).

```
[ ]  X = df_empleados.drop(['Attrition'], axis=1)
     y = df_empleados['Attrition']

     X = pd.get_dummies(X, drop_first=True)

     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

     rf = RandomForestClassifier(n_estimators=100, max_depth=5, random_state=42)

     rf.fit(X_train, y_train)

     y_pred = rf.predict(X_test)

     accuracy = accuracy_score(y_test, y_pred)

     print('Random Forest Classifier accuracy:', accuracy)

     Random Forest Classifier accuracy: 0.8372093023255814
```

El accuracy obtenido del modelo Random Forest Classifier nos indica lo bien que el modelo es capaz de predecir si un empleado abandonará la empresa (Attrition= yes), o se quedará (Attrition= No), basándose en las características dadas. En este caso, el accuracy score indica que el modelo es capaz de predecir correctamente el resultado de aproximadamente el 84% de las muestras de datos de prueba.

EQUILIBRIO DE DATOS MEDIANTE SMOTE Y RANDOM UNDER SAMPLER.

Como hemos dicho antes tenemos un conjunto de datos en el que la variable target está desequilibrada, y eso puede dar lugar a un rendimiento predictivo deficiente, en concreto para la clase minoritaria.

Por ello utilizamos esta técnica para sobre muestreo, con el objetivo equilibrar la distribución de clases, generando ejemplos sintéticos de la clase minoritaria. seleccionamos una clase minoritaria, encontrando sus k vecinos más cercanos y creando nuevas instancias a lo largo de los segmentos de

línea que conectan esos vecinos.

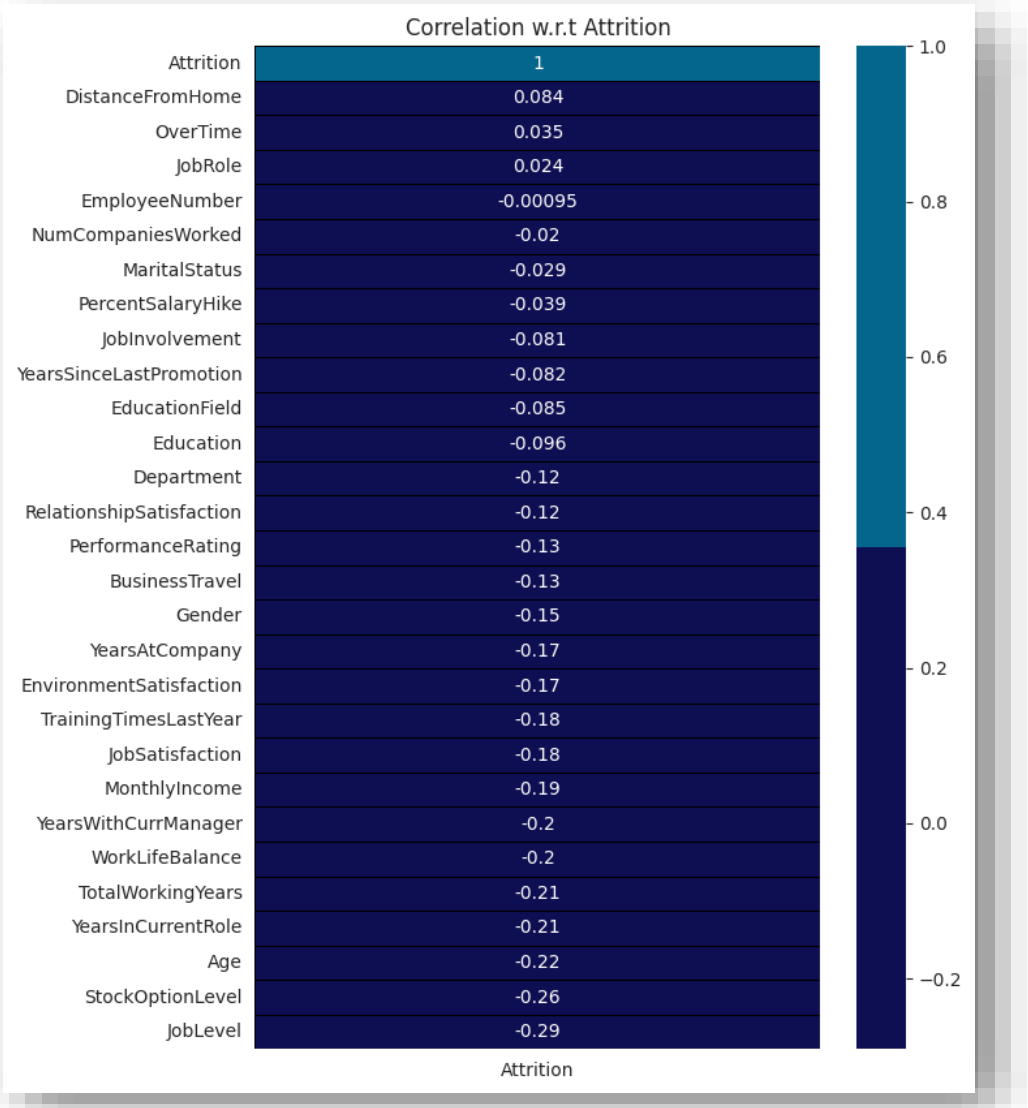
Por otro lado usamos, Random Under Sampler que selecciona aleatoriamente un subconjunto de la clase mayoritaria para equilibrar la distribución (por medio del submuestreo).

Ambas técnicas se utilizaron con el objetivo de tratar el desequilibrio de clases, aunque utilizan enfoques diferentes. SMOTE genero ejemplos para aumentar la clase minoritaria, mientras que Random Under Sampler redujo la clase mayoritaria descartando instancias aleatoriamente.

```
cols = list(df_empleados.columns)
cols.remove('Attrition')

over = SMOTE(sampling_strategy = 0.85)
under = RandomUnderSampler(sampling_strategy = 0.1)
f1 = df_empleados.loc[:,cols]
t1 = df_empleados.loc[:,'Attrition']

steps = [('over', over)]
pipeline = Pipeline(steps=steps)
f1, t1 = pipeline.fit_resample(f1, t1)
Counter(t1)
```

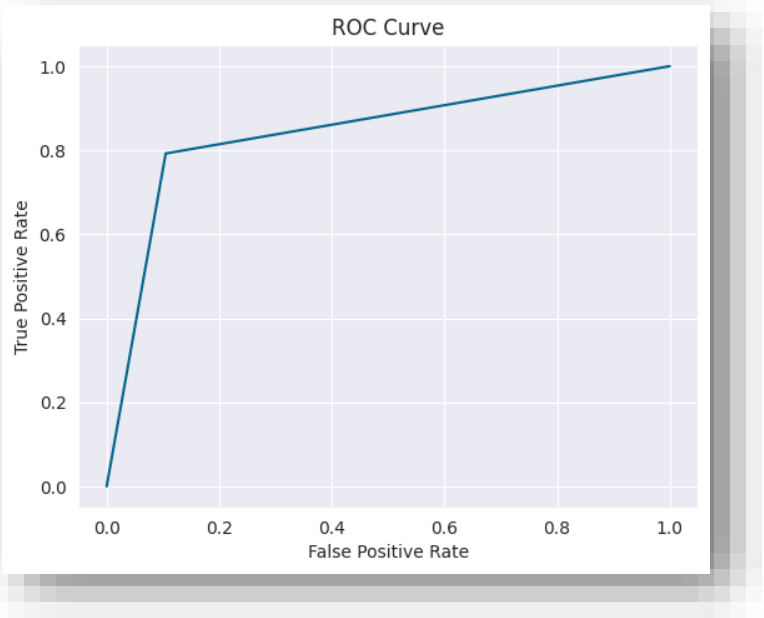
```
Counter({1: 1525, 0: 1795})
```

PARA VISUALIZAR LA MATRIZ DE CORRELACIONES, CREAMOS UN NUEVO MARCO DE DATOS QUE CONTIENE LOS VALORES DE X_TRAIN & Y_TRAIN. DE ESTE MODO, RECHAZAMOS TODO LO QUE ESTÉ FUERA DE LOS DATOS DE ENTRENAMIENTO PARA EVITAR LA FUGA DE DATOS.

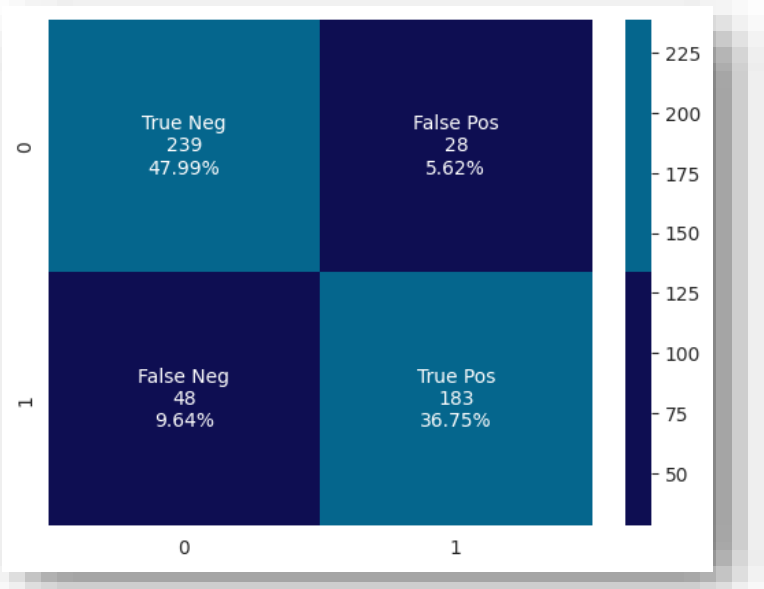


CREAMOS NUEVOS MODELOS.

XGBOOST CLASSIFIER.



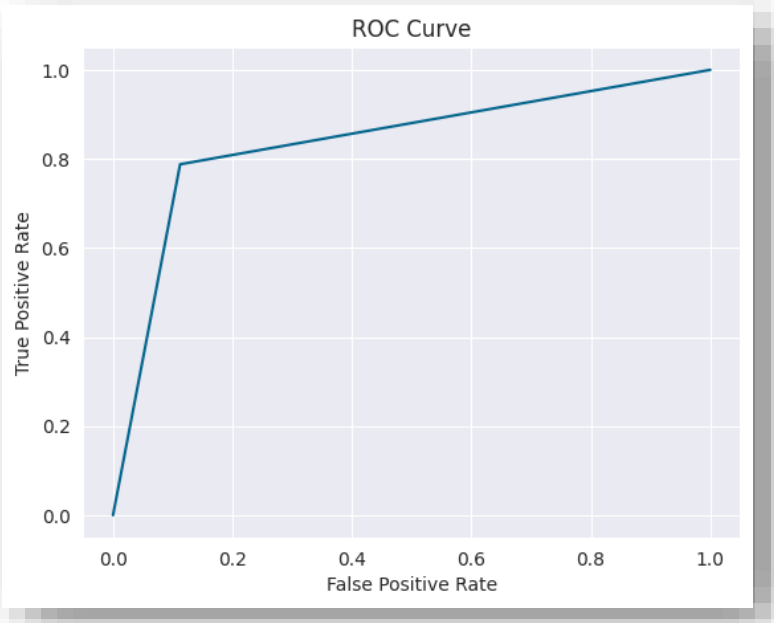

```
AUC: 0.8436694391750571
              precision    recall  f1-score   support

           0       0.83      0.90      0.86       267
           1       0.87      0.79      0.83       231

    accuracy                           0.85       498
   macro avg       0.85      0.84      0.85       498
weighted avg       0.85      0.85      0.85       498
```



XGBoostClassifier.

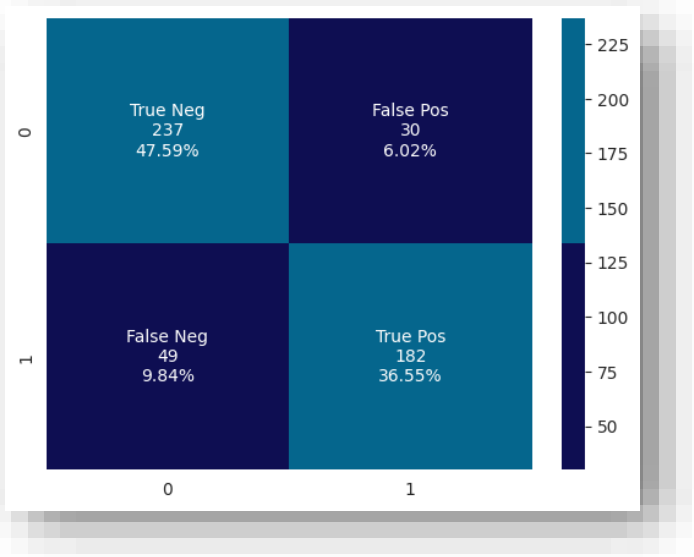
La validación cruzada del 92,41 % indica la precisión promedio del XGBoostClassifier cuando se evalúa mediante la validación cruzada. En este caso, el XGBoostClassifier logró una precisión promedio del 92,41 % en todos los pliegues de validación cruzada. La puntuación de AUC del 84,37 % alude que XGBoostClassifier tiene una buena capacidad para distinguir entre clases positivas y negativas. Estos puntajes indican que XGBoostClassifier está funcionando bien en la tarea de clasificación, logrando una alta precisión y un buen nivel de discriminación entre clases.

LGBM CLASSIFIER.



AUC: 0.837759618658495

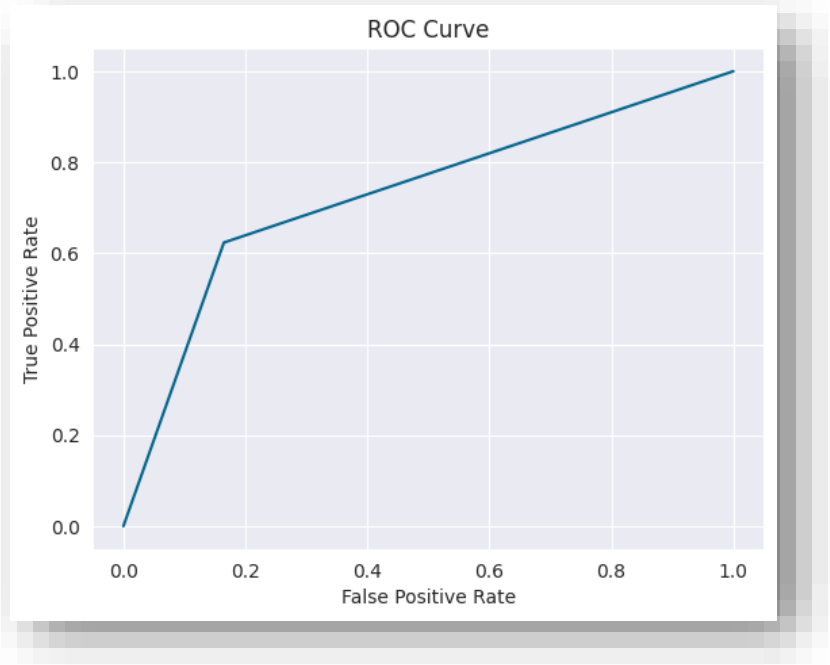
	precision	recall	f1-score	support
0	0.83	0.89	0.86	267
1	0.86	0.79	0.82	231
accuracy			0.84	498
macro avg	0.84	0.84	0.84	498
weighted avg	0.84	0.84	0.84	498



LGBMClassifier.

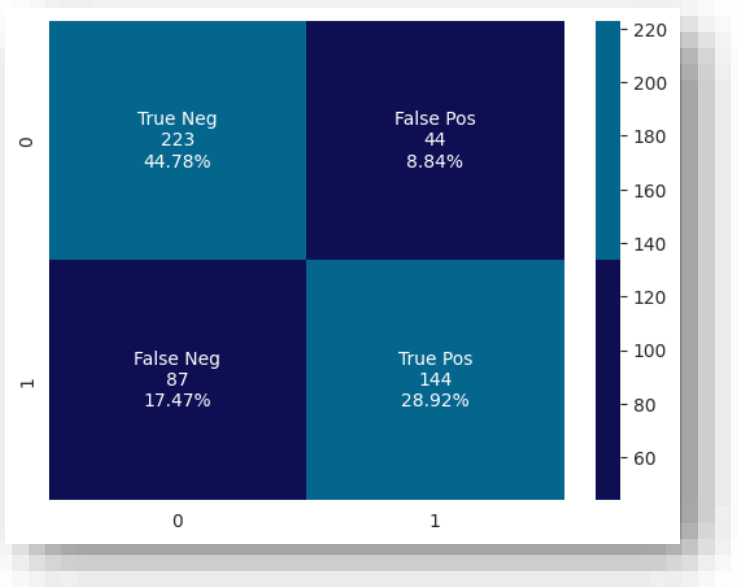
La validación cruzada del 92,70% indica la precisión promedio del LGBMClassifier cuando se evalúa mediante validación cruzada. La puntuación ROC_AUC del 83,78 % indica que LGBMClassifier funciona bien para distinguir entre clases positivas y negativas. En general, tiene una puntuación alta de validación cruzada y una puntuación ROC_AUC relativamente buena, el LGBMClassifier parece ser bueno para las tareas de clasificación.

DECISION TREE CLASSIFIER.

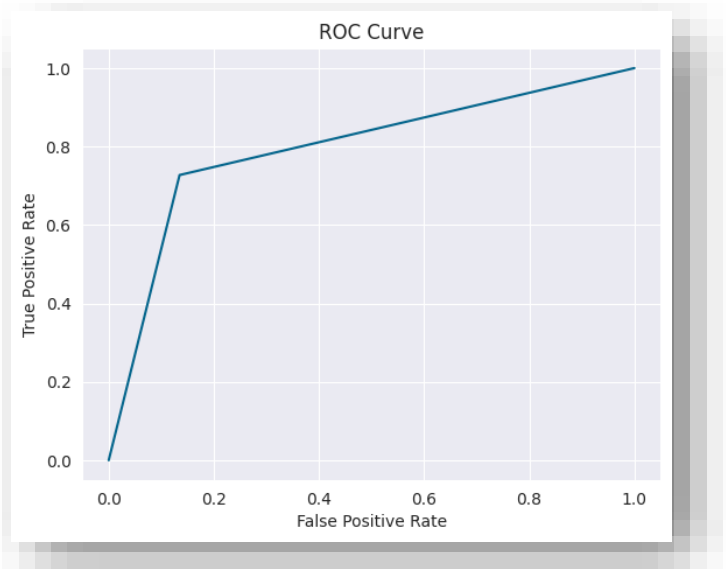


AUC: 0.7292913079429933

	precision	recall	f1-score	support
0	0.72	0.84	0.77	267
1	0.77	0.62	0.69	231
accuracy			0.74	498
macro avg	0.74	0.73	0.73	498
weighted avg	0.74	0.74	0.73	498

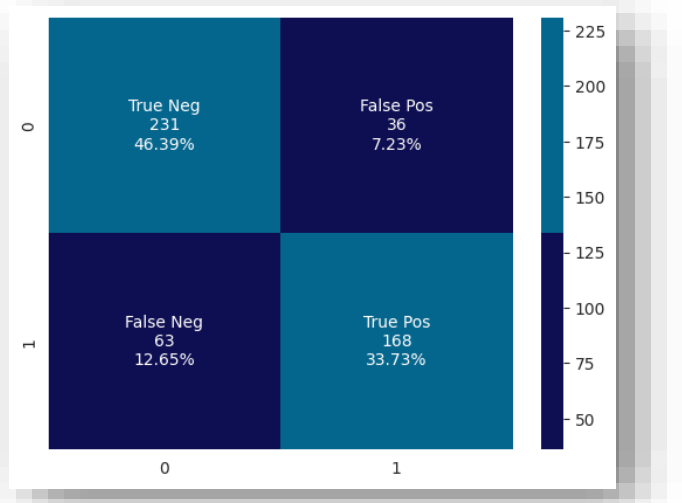


RANDOMFOREST CLASSIFIER



AUC: 0.796220633299285

	precision	recall	f1-score	support
0	0.79	0.87	0.82	267
1	0.82	0.73	0.77	231
accuracy			0.80	498
macro avg	0.80	0.80	0.80	498
weighted avg	0.80	0.80	0.80	498



RandomForest Classifier.

La validación cruzada del 86,80 % sugiere que el modelo funciona bien con los datos con los que se entrenó, ósea el modelo logró una precisión relativamente alta durante este proceso de evaluación. Por otro lado, una puntuación ROC_AUC del 77,99% sugiere que su modelo tiene un poder discriminatorio razonablemente bueno, aunque no es tan alto como la puntuación de validación cruzada. Esto indica que el modelo funciona bien en términos de precisión pero puede no tener el mismo nivel de rendimiento cuando se trata de distinguir entre las clases positivas y negativas.

FUTURAS LÍNEAS.

Hemos demostrado que es un conjunto de datos desequilibrado, ya que hay muy pocos ejemplos de la clase “Attrition Yes” para que el modelo aprenda efectivamente el límite de decisión, aplicamos SMOTE. Lo que nos permitió llevar a un mejor rendimiento de los modelos.

Pero nuestro Proyecto seguirá dinámico, probaremos diferentes técnicas: weighting (promedio ponderado), down-sampling (eliminar aleatoriamente casos de la clase mayoritaria) y up-sampling (replicar aleatoriamente instancias en la clase minoritaria).

CONCLUSIÓN.

Descubrimos factores que afectan a la deserción de los empleados, y así a continuación, tomar medidas para reducir esta tasa.

Construimos un modelo, basado en los factores de los empleados, para predecir si es probable que ese empleado se desvincule o no.

A continuación, se describen las principales observaciones:

- Los hombres tienen un mayor índice de abandono.
- Las mujeres ganan un poco más que los hombres.
- Un empleado con 6 años en su puesto actual gana más que uno con 14 años en su puesto actual.
- Los profesionales jóvenes son más propensos a dejar la empresa.
- Los representantes de ventas tienden a abandonar más otros, porque sus ingresos son inferiores.
- A medida que aumenta el nivel de estudios, aumentan los ingresos medios mensuales.
- Los empleados tienden a marcharse más cuando la distancia al domicilio es superior a 10 km.
- Los empleados solteros tienden a marcharse porque tienen ingresos mensuales más bajos que los demás.
- Es necesario implantar una estructura innovadora para los empleados con 1 año de experiencia, ya que contribuye en gran medida al porcentaje de abandono.
- Es necesario idear mejores opciones sobre acciones para las personas con más de 6 años en su puesto actual, ya que el desgaste parece aumentar gradualmente con una caída de los ingresos mensuales.
- Los empleados insatisfechos con las condiciones del entorno y la JobSatisfaction tienden a marcharse más en comparación con los demás.

PARA PREVENIR EL DESGASTE, SE SUGIERE.

- Retribuir equitativamente a los empleados con el mismo nivel de trabajo, la misma implicación en el trabajo y la misma función, con ingresos mensuales y acciones casi iguales.
- Aumentar en sueldos de a aquellos empleados hagan horas extras, y asegurarse de que a los que las hacen se les pague más que a los que no las hacen.
- La formación debe impartirse por niveles, y el número de sesiones debe depender de las necesidades específicas de cada departamento.
- Los directivos deben recibir formación sobre los requisitos de su función, así como sobre la forma de impartir una formación eficaz a los miembros de su equipo.

- Se pudo ver fuertes características que podrían determinar el desgaste de los empleados, tales como (Horas extras, Distancia de casa, Años desde la última promoción), se debería empezar trabajando en ello.