



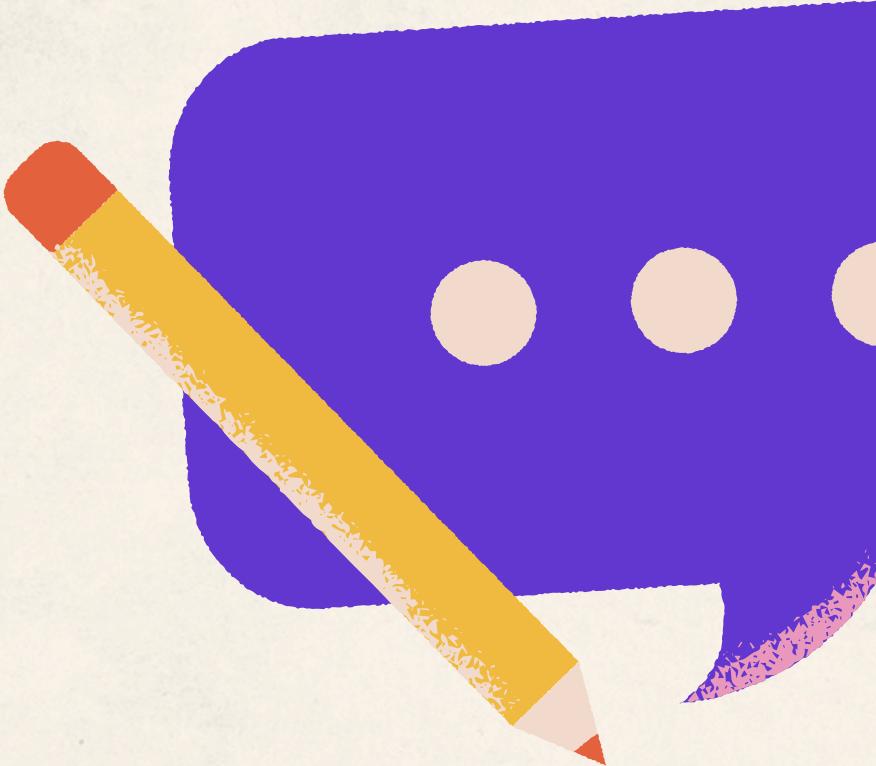
OPTIMIZACIÓN DE TALENTO



ABC Corporation

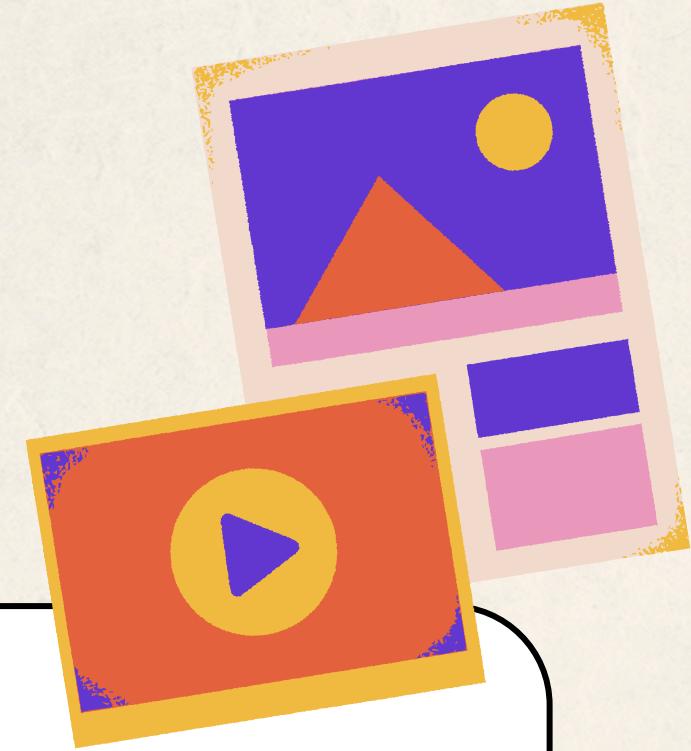


by Data CIRCUS



ÍNDICE

- ✓ Nuestro Equipo
- ✓ Fases del proceso
- ✓ Misión
- ✓ Reporte de resultados
- ✓ Proceso
- ✓ Debilidades y fortalezas
- ✓ Herramientas



Nuestro Equipo



Belén

Scrum Master
*Data Analyst,
Lead Engineer &
BBDD Engineer*



Viviana

**BBDD Engineer -
Visualization
Specialist &
Presentation
Designer**



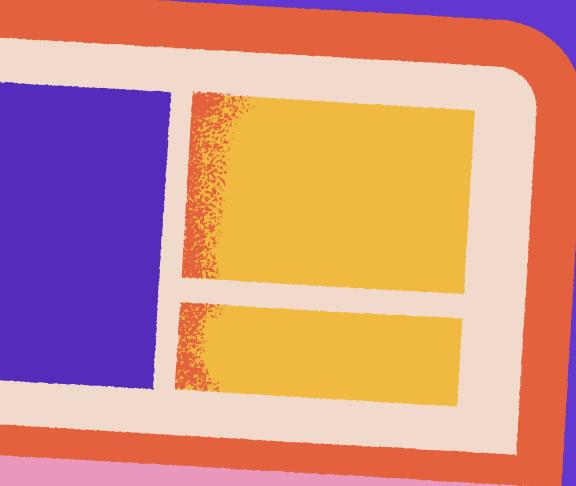
Cristina

**Data Engineer
(Support) - A/B
Testing Specialist &
Presentation
Designer**



Gloria

**ETL Developer (Junior
Support) &
Visualization
(Support)**





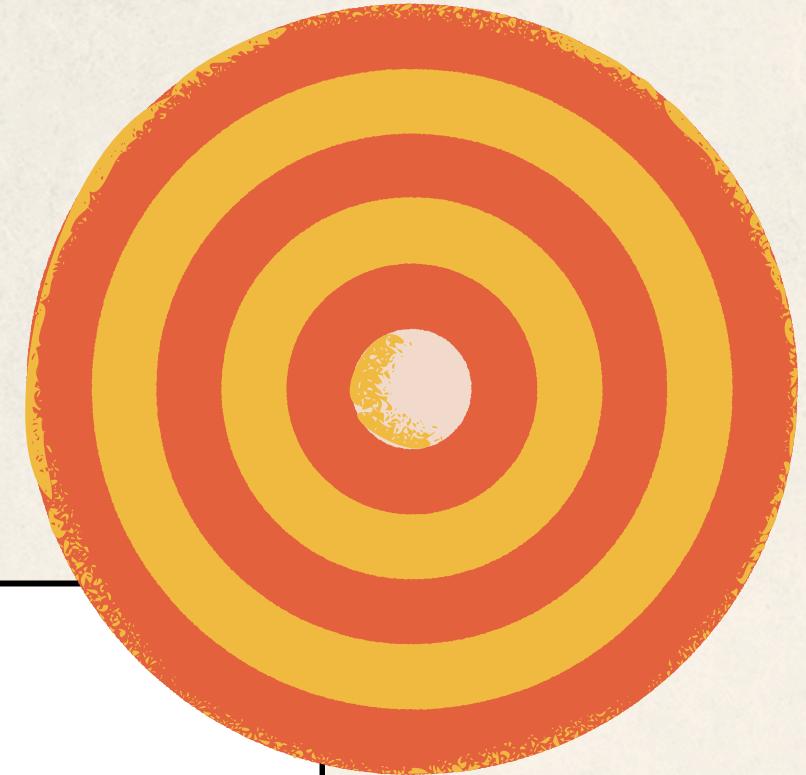
Misión

Nuestra misión es identificar factores clave que influyen en la satisfacción laboral y, en última instancia, en la retención de empleados. Para ello, hemos llevado a cabo un complejo proceso de análisis de datos que incluye: proceso EDA, transformación de datos, Testing A/B, visualizaciones, creación de una base de datos MySQL y proceso ETL.



Proceso

- Fase 1: Análisis Exploratorio de Datos(EDA).
- Fase 2: Transformación de los datos
- Fase 3: Diseño de BBDD e Inserción de los Datos (estructura).
- Fase 4: Problema de A/B Testing.
- Fase 5: Creación de una ETL.
- Fase 6: Reporte de los resultados.



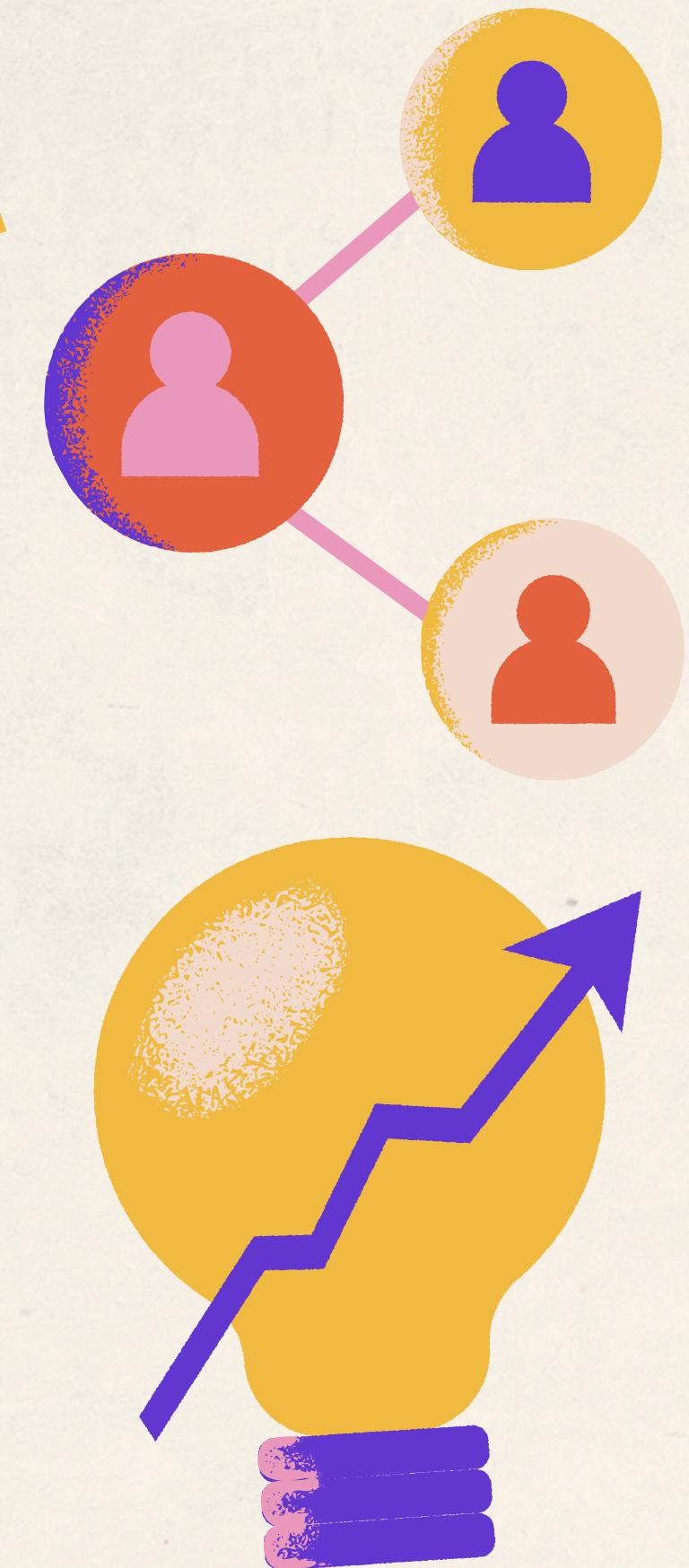
Herramientas

Librerías

- pandas
- numpy
- matplotlib
- seaborn
- scikit-learn
- mysql connector
- scipy stats
- chi2_contingency

Tecnologías

- Operating system: Windows 10 Home
- Development Environment: Jupyter Notebook, Visual Studio Code
- Programming Language: Python
- Libraries specified above
- Version Control: Git, GitHub
- Dependency Management: Pip
- MySQL Workbench



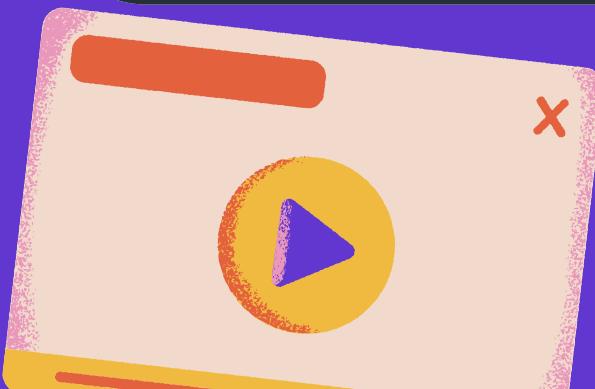
Análisis Exploratorio de Datos (EDA)



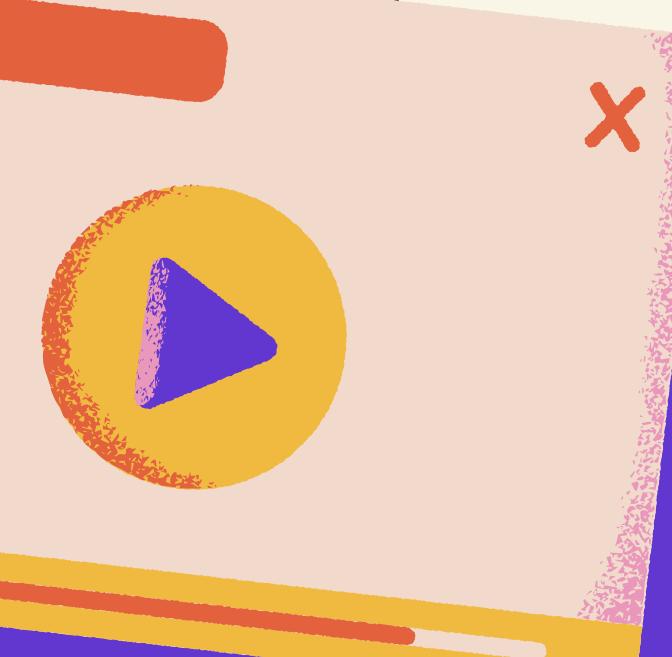
- Durante esta fase hemos hecho un análisis profundo de los datos aportados
- Mediante métodos de pandas hemos obtenido información sobre la estructura de los datos
- Hemos hecho una exploración inicial de los datos para identificar potenciales problemas (valores nulos y duplicados, valores raros o datos faltantes)

```
# Review the main statistical data of the DataFrame  
df_employees.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	1614.0	806.500000	466.065982	0.0	403.25	806.5	1209.75	1613.0
DistanceFromHome	1614.0	4.527261	14.591913	-49.0	2.00	5.0	11.00	29.0
Education	1614.0	2.925031	1.022357	1.0	2.00	3.0	4.00	5.0
employeecount	1614.0	1.000000	0.000000	1.0	1.00	1.0	1.00	1.0
EnvironmentSatisfaction	1614.0	4.294919	6.993559	1.0	2.00	3.0	4.00	49.0
Gender	1614.0	0.398389	0.489718	0.0	0.00	0.0	1.00	1.0
JobInvolvement	1614.0	2.739777	0.711567	1.0	2.00	3.0	3.00	4.0
JobLevel	1614.0	2.068154	1.101344	1.0	1.00	2.0	3.00	5.0
JobSatisfaction	1614.0	2.738538	1.106163	1.0	2.00	3.0	4.00	4.0
MonthlyRate	1614.0	14284.495663	7110.414585	2094.0	8001.00	14248.5	20364.00	26999.0
NUMCOMPANIESWORKED	1614.0	2.673482	2.506152	0.0	1.00	2.0	4.00	9.0
PercentSalaryHike	1614.0	15.165428	3.648610	11.0	12.00	14.0	18.00	25.0
RelationshipSatisfaction	1614.0	2.704461	1.079031	1.0	2.00	3.0	4.00	4.0
StockOptionLevel	1614.0	0.791202	0.842396	0.0	0.00	1.0	1.00	3.0
TrainingTimesLastYear	1614.0	2.809789	1.297765	0.0	2.00	3.0	3.00	6.0
YearsAtCompany	1614.0	7.132590	6.124237	0.0	3.00	5.0	9.00	40.0
YearsSinceLastPromotion	1614.0	2.245973	3.235665	0.0	0.00	1.0	3.00	15.0
YEARSWITHCURRMANAGER	1614.0	4.220570	3.562695	0.0	2.00	3.0	7.00	17.0
DateBirth	1614.0	1986.076208	9.101332	1963.0	1980.00	1987.0	1993.00	2005.0
NUMBERCHILDREN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN



Transformación de los datos





Tras analizar la base de datos, se procede a la transformación de los datos con el objetivo de tener una base de datos de mejor calidad

- **1a Fase:**

- 1) Modificación tipo de dato de las columnas
- 2) Eliminación columnas innecesarias

- **2a Fase: Homogenización de los datos**

- 1) Homogenización de los títulos de las columnas
- 2) Homogenización de los datos de las columnas

- **3a Fase: Transformación de los datos**

- 1) Números negativos
- 2) Identificación de outliers
- 3) Gestión de nulos*
- 4) Gestión de duplicados*





DUPLICADOS

- Nos encontramos con 59 líneas duplicadas y 534 valores duplicados
- Gracias a la entrevista con nuestro cliente podemos tomar ciertas decisiones para eliminar los duplicados
- Debido a la alta cantidad de filas duplicadas por la repetición de "EmployeeNumber", realizamos un filtro para revisar las filas donde esta columna tuviera un valor numérico. Realizamos las gestiones de limpieza según conversación con el cliente y, por último, concatenamos los resultados con las filas en las que "EmployeeNumber" era Unknown.

```
[96] df_clean_final['EmployeeNumber'].duplicated().sum()
...
430

# Check for duplicate rows after the elimination
df_clean_final.duplicated().sum()

[97]
...
21

> ▾
[98] df_clean_final['EmployeeNumber'].value_counts()
...
EmployeeNumber
Unknown    431
1833        1
1683        1
1694        1
1702        1
...
785         1
861         1
864         1
878         1
178         1
Name: count, Length: 1080, dtype: int64
```





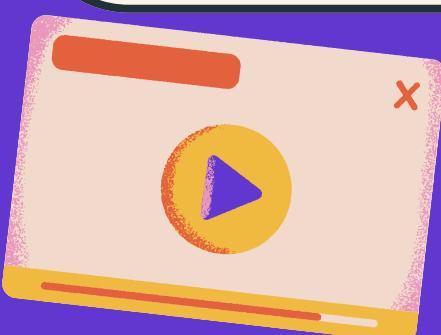
GESTIÓN NULOS

- Revisamos los valores nulos de aquellas columnas que tienen nulos. Hay columnas que tienen un gran porcentaje de nulos como Department: 1312, StandardHours: 1195 o Yearsincurrentrole:1580
- Por ejemplo: “EmployeeNumber” tiene 431 valores nulos, lo que es un 26%. Precisamente, por la relevancia de “EmployeeNumber” decidimos reemplazar los valores nulos por Unknown.
- En otras columnas, decidimos reemplazar los nulos por la moda como en “BusinessTravel” y en otras aplicamos métodos avanzados como en “MonthlySalary”, aplicando el IterativeImputer

```
percent_null = (df_clean.isnull().sum()/df_clean.shape[0])*100  
percent_null
```

```
EmployeeNumber      0.0  
Age                0.0  
Attrition          0.0  
BusinessTravel      0.0  
Department         0.0  
DistanceFromHome   0.0  
Education           0.0  
EducationField      0.0  
EnvironmentSatisfaction 0.0  
Gender              0.0  
JobInvolvement     0.0  
JobLevel            0.0  
JobRole             0.0  
JobSatisfaction    0.0  
MaritalStatus       0.0  
MonthlyIncome       0.0  
MonthlyRate         0.0  
NumCompaniesWorked 0.0  
OverTime            0.0  
PercentSalaryHike   0.0  
PerformanceRating   0.0  
RelationshipSatisfaction 0.0  
StockOptionLevel    0.0  
TotalWorkingYears   0.0  
TrainingTimesLastYear 0.0  
...  
YearsSinceLastPromotion 0.0  
YearsWithCurrManager 0.0  
DateBirth           0.0  
RemoteWork          0.0  
dtypes: float64
```

Python



Diseño de BBDD e Inserción de los Datos (estructura)

Creación de una BDD



1) Diseño de la Estructura de la BBDD

2) Creación de la BBDD

3) Inserción de Datos Iniciales

```
-- Schema ABC_CORP_EMPLOYEES
-- 
-- Schema ABC_CORP_EMPLOYEES
-- 
CREATE SCHEMA IF NOT EXISTS `ABC_CORP_EMPLOYEES` DEFAULT CHARACTER SET utf8 ;
USE `ABC_CORP_EMPLOYEES` ;

-- Table `ABC_CORP_EMPLOYEES`.`PersonalDetails`
-- 
CREATE TABLE IF NOT EXISTS `ABC_CORP_EMPLOYEES`.`PersonalDetails` (
  `EmployeeNumber` INT NOT NULL AUTO_INCREMENT,
  `Age` INT NULL DEFAULT NULL,
  `Education` INT NOT NULL,
  `EducationField` VARCHAR(45) NULL DEFAULT NULL,
  `Gender` VARCHAR(45) NULL DEFAULT NULL,
  `MaritalStatus` VARCHAR(45) NULL DEFAULT NULL,
  `DateBirth` INT NULL DEFAULT NULL,
  PRIMARY KEY (`EmployeeNumber`)
);

-- Table `ABC_CORP_EMPLOYEES`.`CompensationDetails`
```

A/B Testing



A/B testing

El objetivo de esta fase es determinar si existe una relación entre el nivel de satisfacción en el trabajo y la rotación de empleados, y si es así, cuál es la magnitud de esa relación.

Nivel de Satisfacción

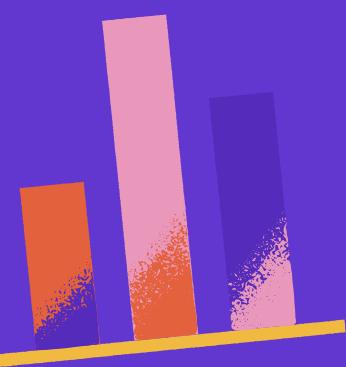
- Empleados Grupo A: Nivel de satisfacción en el trabajo =>3
- Empleados Grupo B: Nivel de satisfacción en el trabajo <3

Attrition

- Nos indica si los empleados han dejado o no la empresa

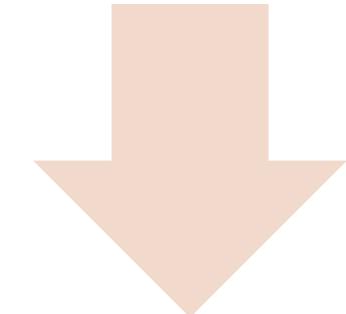
Tasa de rotación = Empleados que han dejado la empresa / Nivel de Satisfacción

Conclusiones A/B testing

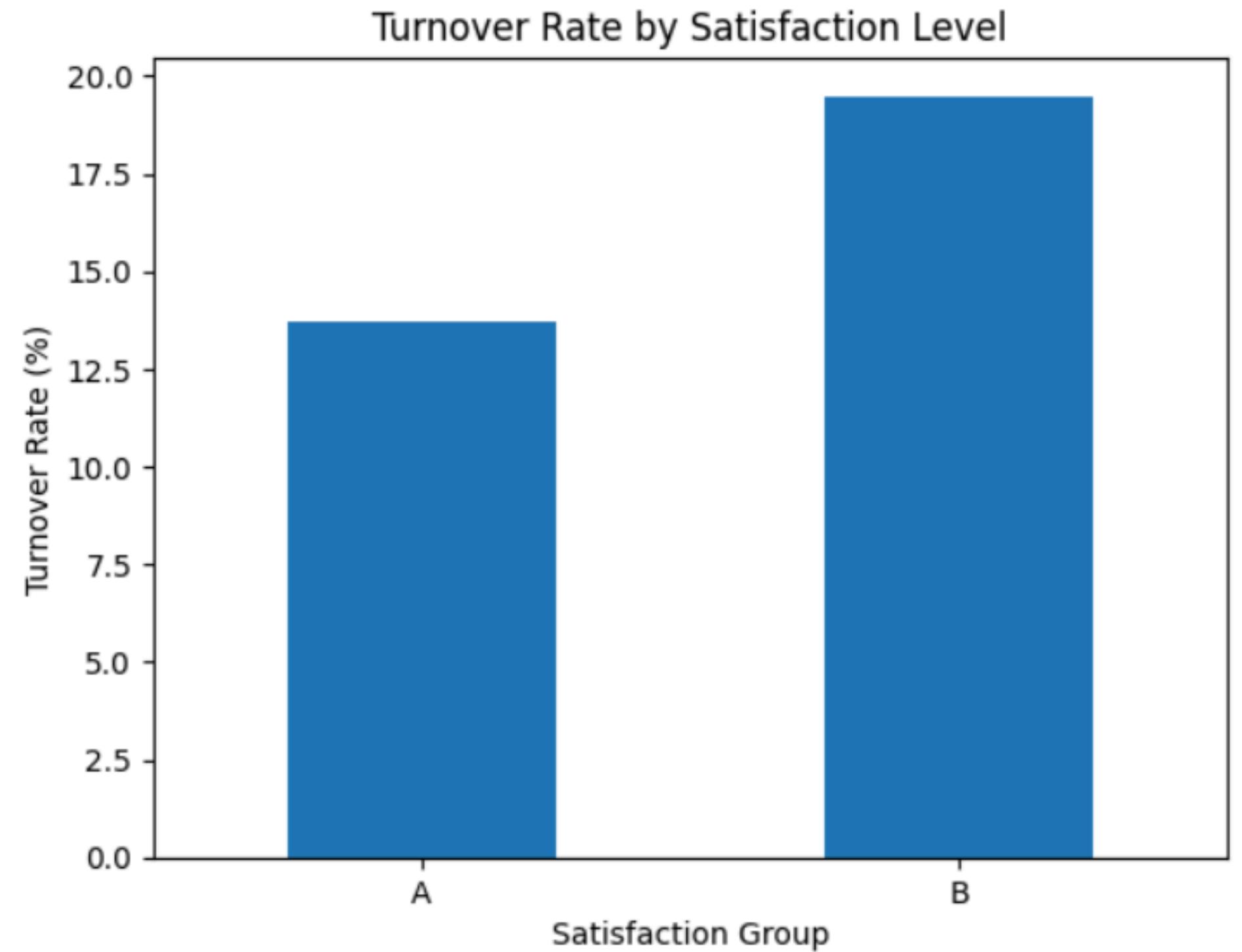


Relación significativa entre el nivel de satisfacción y la rotación de empleados

Con base en estos resultados, se puede concluir que los empleados con mayor satisfacción laboral tienden a permanecer más tiempo en la empresa, mientras que aquellos con menor satisfacción tienen más probabilidades de irse.



Mejorar los factores clave relacionados con la satisfacción podría, reducir la rotación y mejorar la retención del talento.



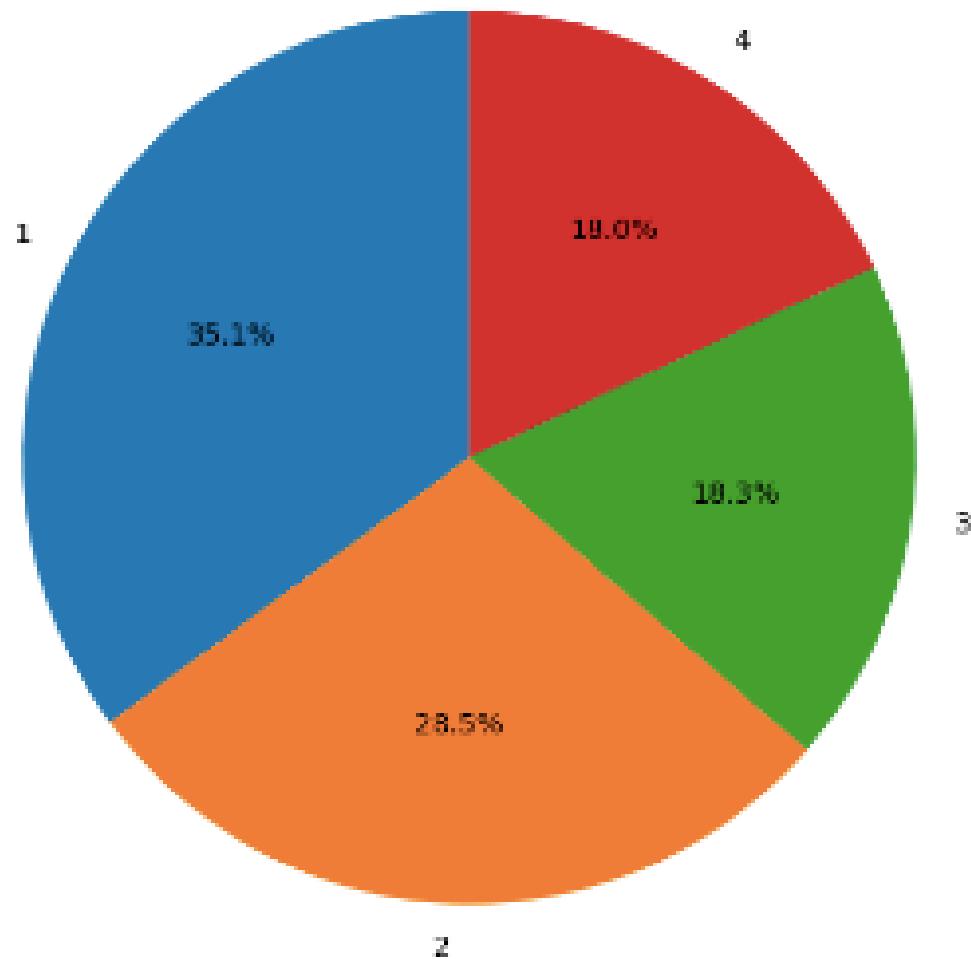
Reporte de los resultados



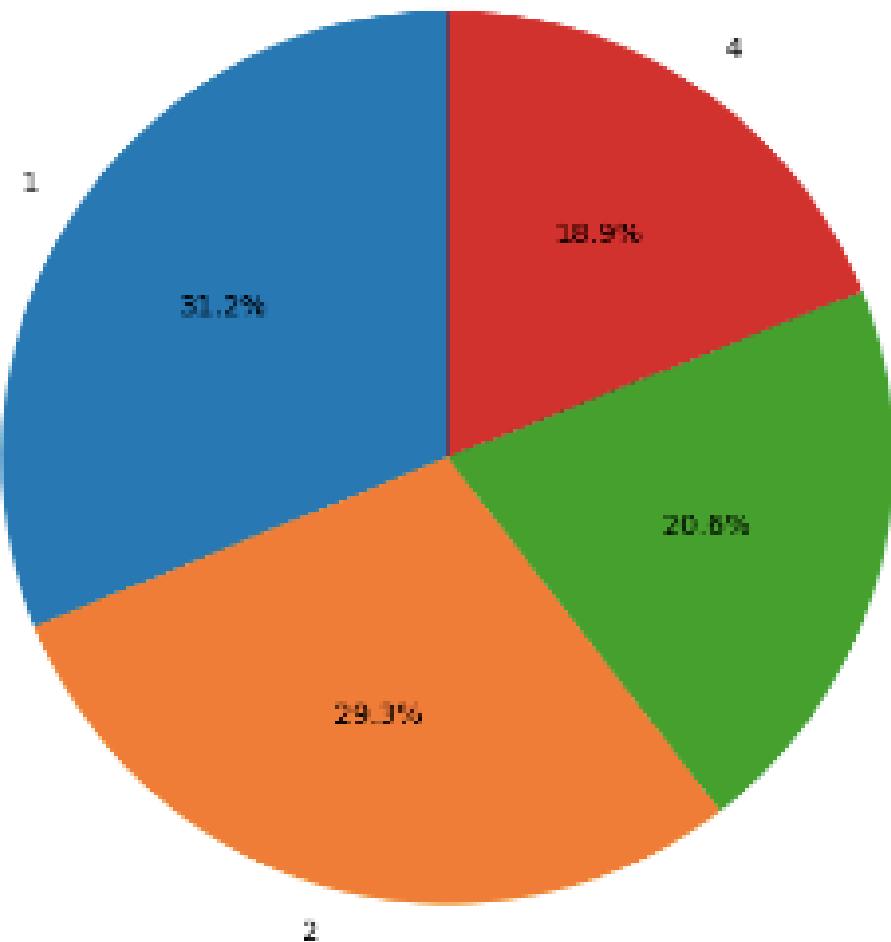
Results



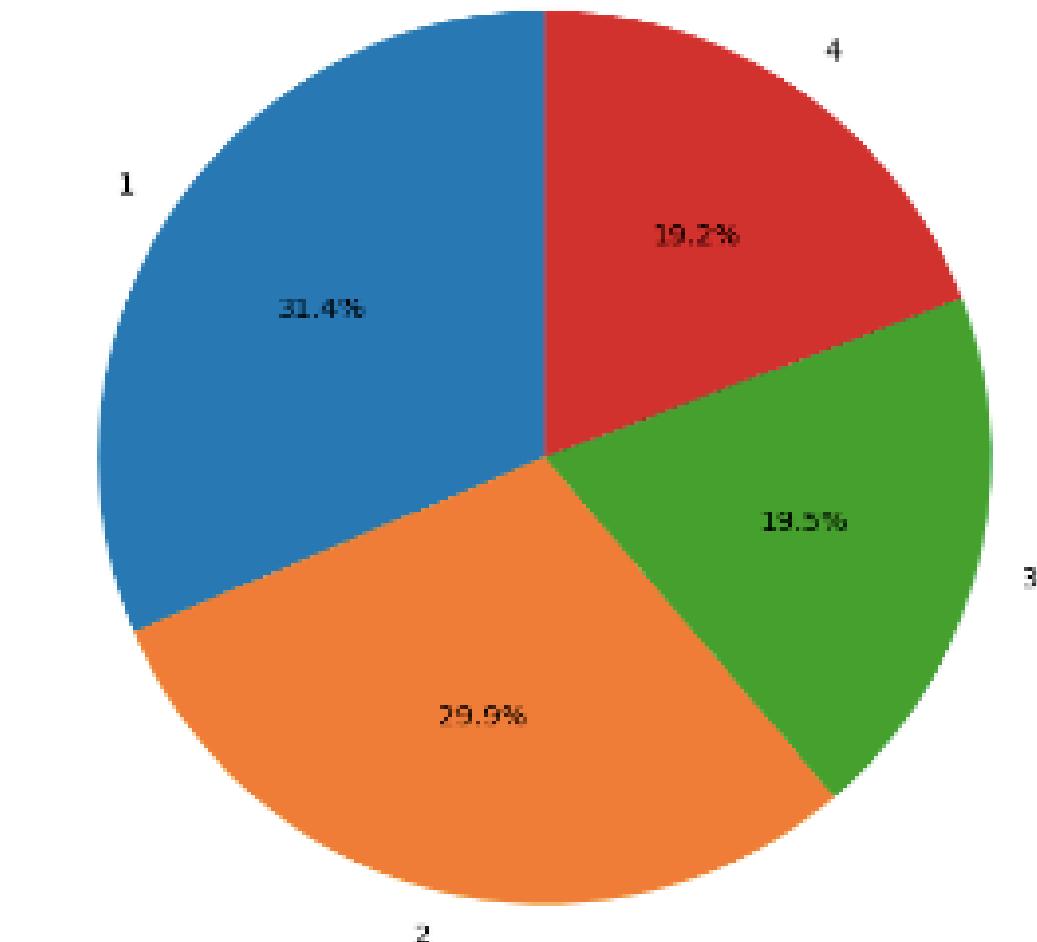
Environment Satisfaction



Relationship Satisfaction



Job Satisfaction



Environment Satisfaction

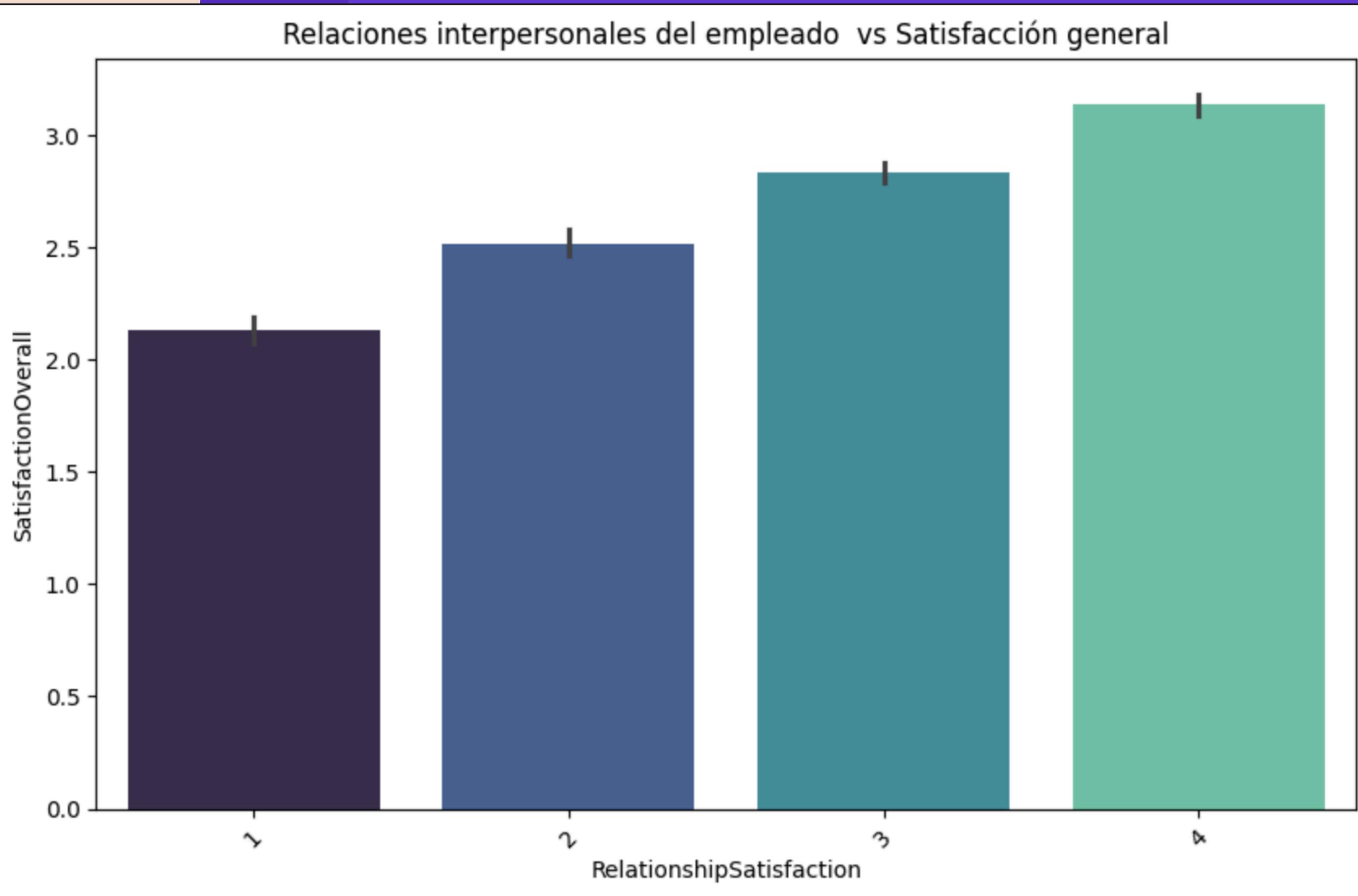
Relationship Satisfaction

Job Satisfaction

Results



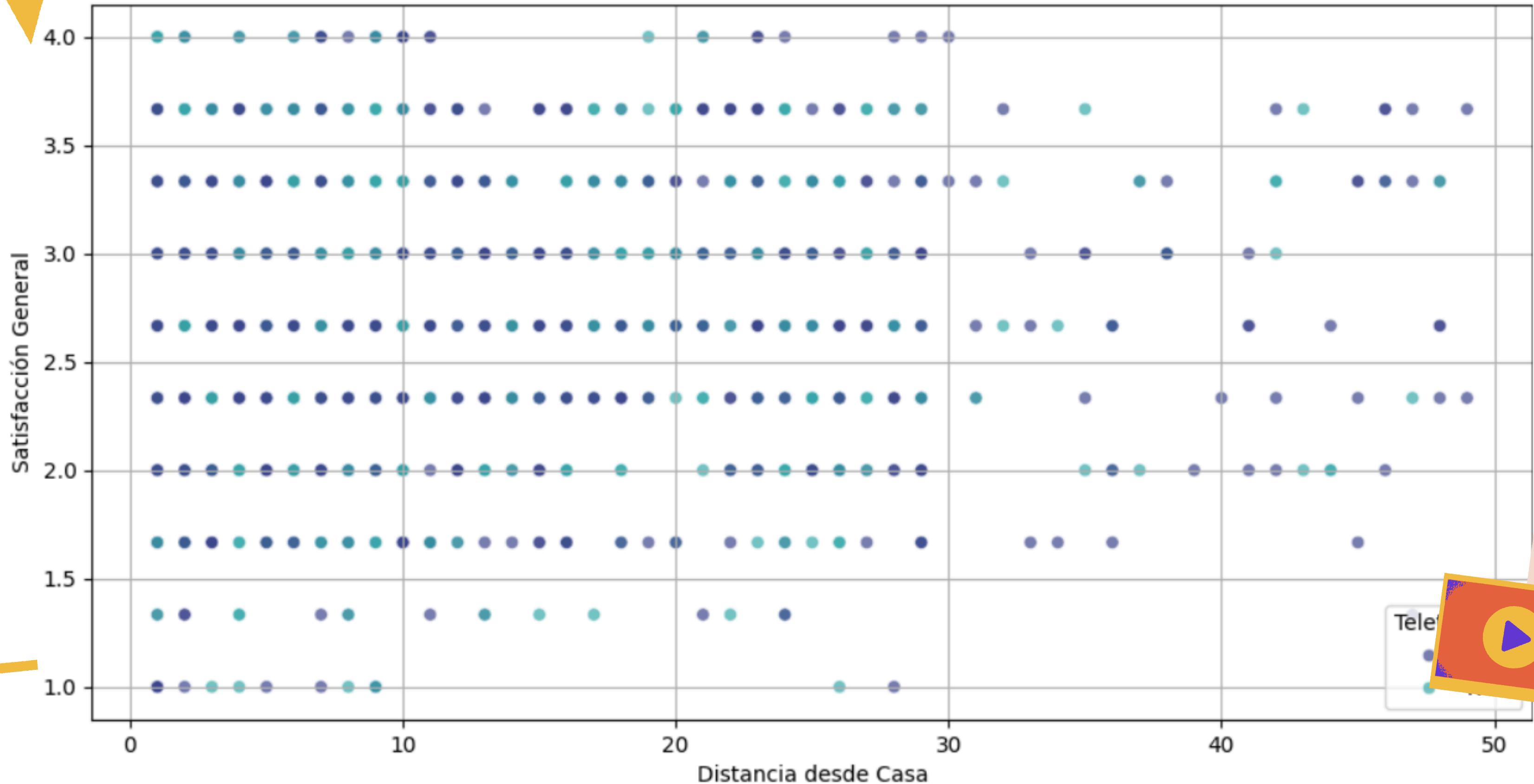
Relaciones interpersonales del empleado vs Satisfacción general



Results



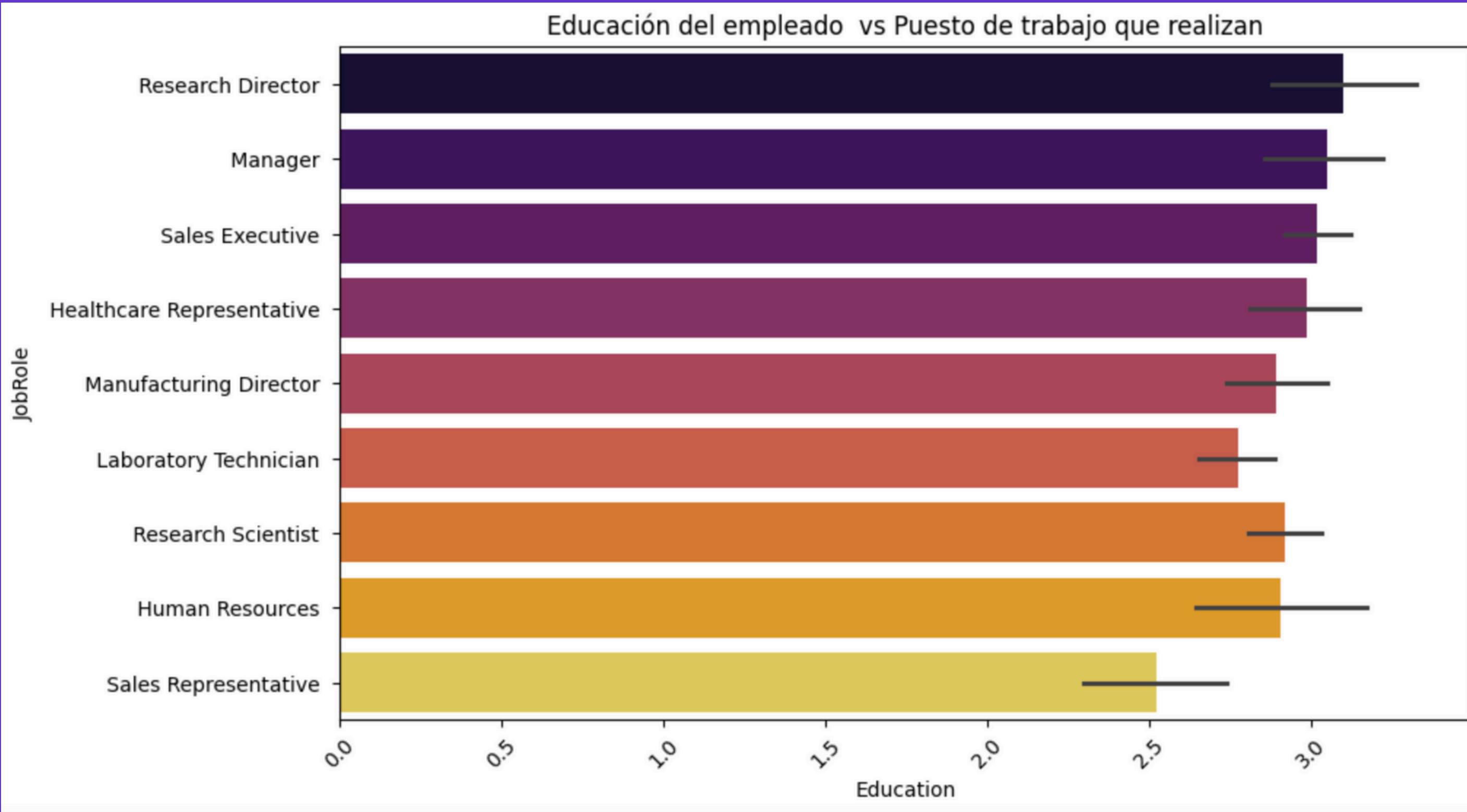
Relación entre Satisfacción General y Distancia desde Casa



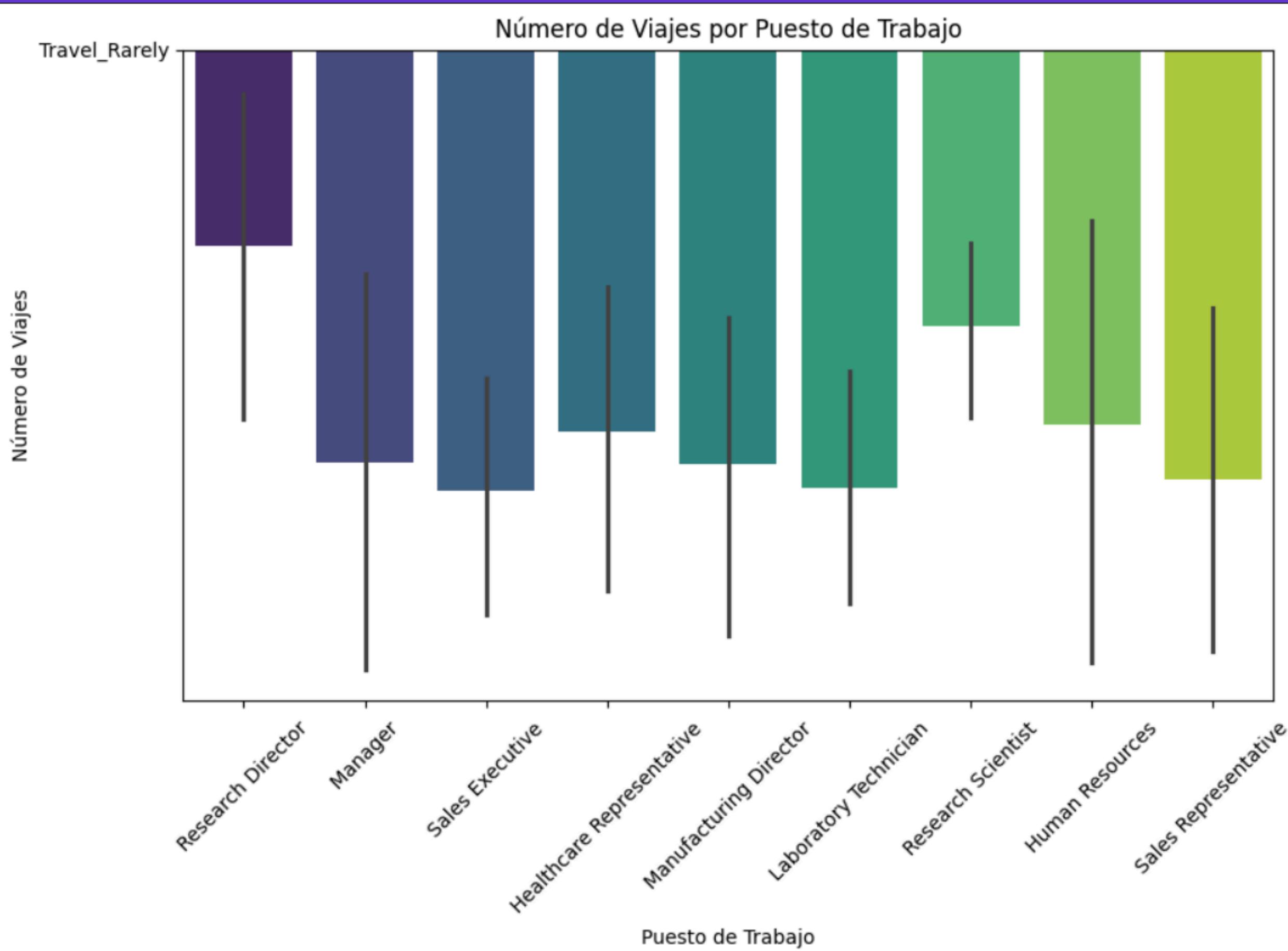
Results



Educación del empleado vs Puesto de trabajo que realizan



Results

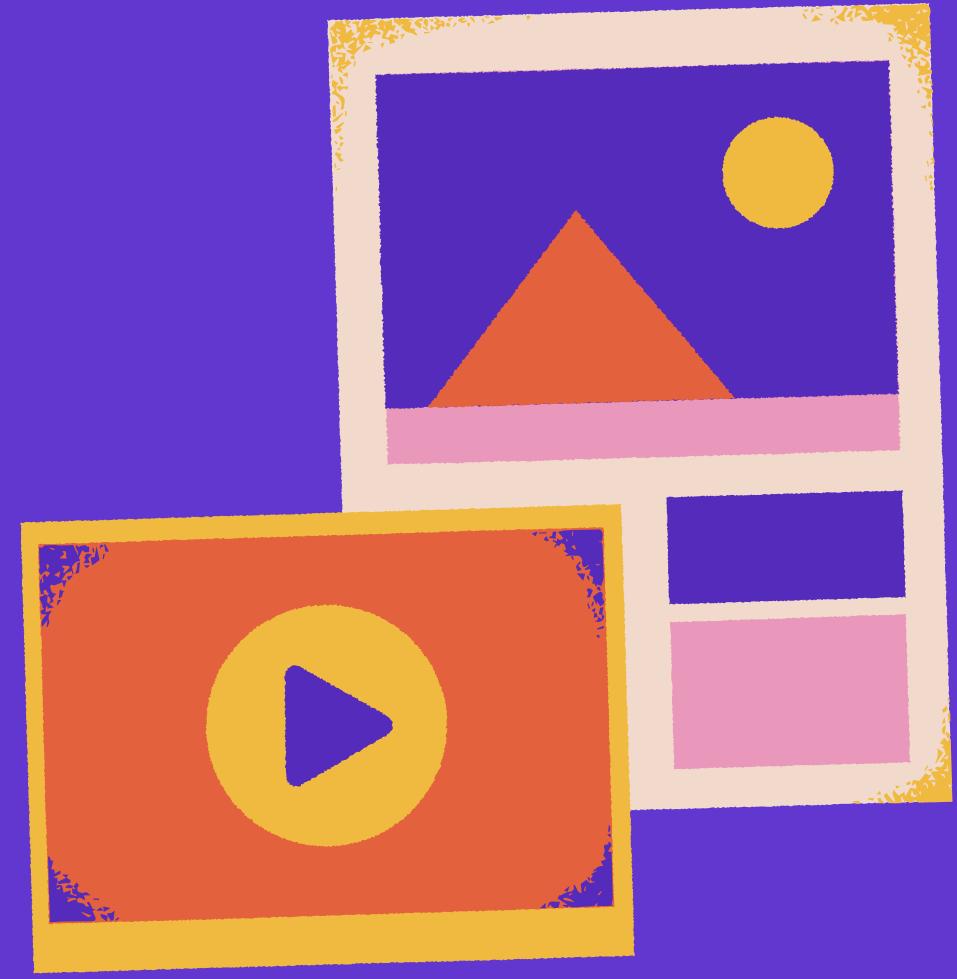
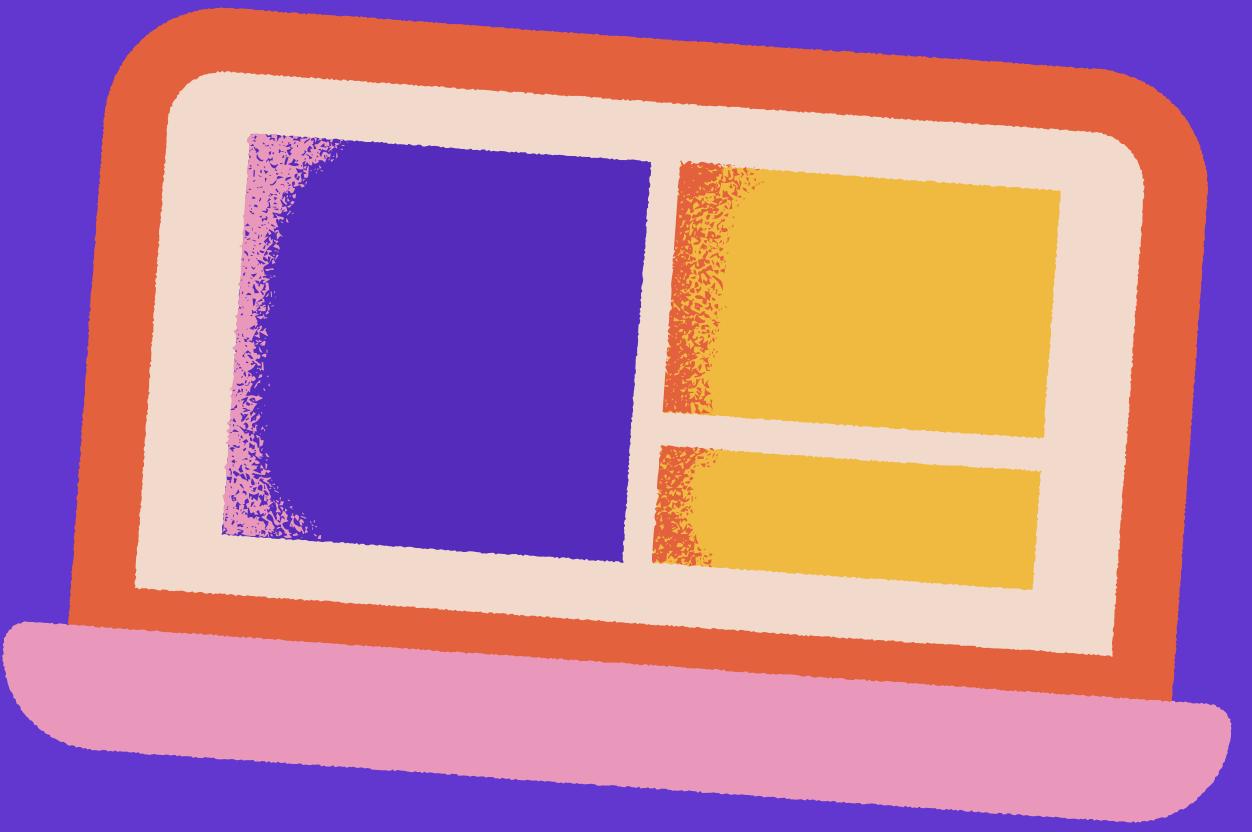




Conclusiones

- **Empleado insatisfecho, tiende a irse de la empresa -> CLAVE**
- Mejora de los datos en la BBDD -> Disminuir los nulos, disminuir los datos duplicados, variables que no indican información relevante
- Desempeño del empleado obtenido en la BBDD con información insuficiente. Mejorar las encuestas internas para obtener un dato calidad más certero
- Cuánto menos distancia a casa, mayor satisfacción
- Una mejora de las relaciones interpersonales del empleado, mejora la satisfacción general del empleado -> Realizar actividades grupales para fomentarlo
- No hay discriminación en la empresa, dado que el género, el número de hijos y el estado civil no influye significativamente (ni en la rotación ni en la satisfacción)





Thank you!!!!