Análisis sobre la Actualidad Laboral en el Sector de IT en Argentina

Entrega Final

Alumno: Gulfo, Maximiliano

Profesor: Russo Locati, Ignacio

Tutor: Benitez, Gustavo

Curso: CoderHouse - Data Science

Comisión: #52295

Fecha de Inicio: 14/10/2023
Fecha de Fin: 04/05/2024

Contenido

Ilustraciones

Introducción

En el sector de Tecnologías de la Información (IT) de Argentina, el análisis de los salarios desempeña un papel crucial para comprender las tendencias laborales y establecer estrategias efectivas de retención de talento.

En este análisis, se explorarán diversas hipótesis relacionadas con los factores que pueden influir en los salarios de los profesionales de IT en Argentina. Desde la relación entre la experiencia laboral y los salarios, hasta la posible brecha salarial por género. El objetivo es descubrir patrones y tendencias significativas que puedan ayudar a las empresas y profesionales a tomar decisiones informadas.

Contexto y Alcance

Contexto

Contexto Comercial:

Este trabajo comenzó a desarrollarse a pedido de distintas Startups del sector tecnológico. El principal desafío es determinar una estructura de salarios para identificar áreas donde se pueden hacer ajustes sin perder talento clave e invertir en concordancia a las tendencias laborales, logrando estar a la vanguardia constantemente. De esta manera, la oferta salarial y la búsqueda será mucho más precisa y eficiente, y se ahorraría tiempo y dinero.

Así mismo, desde distintas organizaciones gubernamentales han pedido informes para poder comprender mejor la situación del mercado laboral tecnológico y tomar decisiones relacionadas con políticas de empleo y educación. Esto colaborará con la mejor distribución del presupuesto público destinado a educación, financiamientos a pymes y la realización de normativa acorde, logrando un sector laboral más equitativo con oportunidades laborales y salarios mejor estructurados.

Problema Comercial:

En base a lo pedido por las Startups se realizará el siguiente trabajo:

- Predecir los salarios mensuales de profesionales en el sector IT en función de variables clave como experiencia, educación y habilidades.
- Clasificar si una habilidad específica será demandada, o no, en el futuro.

Por otro lado, por lo expuesto por los distintos organismos del sector gubernamental, se buscará lo siguiente:

- Segmentar la población de profesionales en el sector IT en grupos demográficos similares.
- Clasificar profesionales en grupos de brechas salariales específicas (por ejemplo, alta, media, baja)
-

Contexto Analítico:

Se tiene un dataset detallado de sueldos en el sector IT de Argentina, recopilado en 2022, el cual surge de una encuesta difundida por los distintos entes gubernamentales y las startups interesadas. En base a esto, se buscarán insights valiosos mediante distintos tipos de análisis (descriptivo, gráficos, modelos de machine learning) para poder responder las preguntas específicas y abordar los problemas comerciales.

La estructura general del trabajo tendrá el siguiente orden:

- Preparación de los Datos
- Exploración Inicial y Creación de Visualizaciones (EDA)
- Desarrollo de modelos predictivos
- Segmentación Demográfica (Seguramente por K-Means)
- Análisis de Resultados
- Recomendaciones y Acciones

Alcance

Este informe está destinado a stakeholders del sector de IT. El alcance del mismo es amplio y puede beneficiar a una variedad de audiencias, desde profesionales individuales hasta empresas, instituciones educativas y entidades gubernamentales. Al dirigir la información a estas audiencias específicas, el análisis del dataset puede tener un impacto significativo y contribuir a la toma de decisiones informada en el ámbito laboral tecnológico en Argentina.

Objetivos Generales

El objetivo se centra en un análisis exhaustivo de las principales variables para entender, y poder explicar de manera armónica y eficiente, los datos conseguidos a través de esta encuesta. Así, se podrá tener una perspectiva más amplia y precisa de la actualidad laboral del sector IT.

Hipótesis

A través del análisis y representación de los datos se buscará dar respuesta a las siguientes conjeturas:

- **La experiencia laboral está correlacionada con el salario**:
  Se espera que profesionales con más años de experiencia ganen salarios más altos.
- **La ubicación geográfica influye en los salarios**:
  La hipótesis plantea que existirá una variación significativa en los salarios según la región geográfica en Argentina, en particular, se cree que los mejores salarios se darán en Buenos Aires y CABA debido a su mayor densidad poblacional y a ser el centro de La Nación.

- * **La educación afecta los niveles salariales**:
  La suposición es que aquellos con títulos más avanzados o certificaciones adicionales ganarán salarios más altos.
  Hipótesis rechazada por falta de datos - Desarrollado y demostrado en sección 6 y 7.
- **La cantidad de empleados y brecha salarial por género:**
  Se espera que el mayor caudal de empleados sea masculino y que exista una brecha salarial a favor de los mismos en el sector de IT.
- **El lenguaje de programación tiene relación con el salario**:
  Las presunciones indican que los lenguajes más populares en la actualidad, como python, tendrán un salario promedio, mientras que lenguajes emergentes, o en pleno crecimiento, tendrán salario más alto debido a la escasez de profesionales aptos para realizar el trabajo.

Objetivo Principal

Trabajar en los distintos modelos ayudará a comprender mejor la dinámica del mercado laboral e identificar distintos patrones y tendencias. De esta manera, tomar decisiones informadas para mejorar el desarrollo profesional de los individuos y las empresas.

A partir de esto, se decide estudiar y realizar los siguientes análisis:

- **Análisis Descriptivo:**
  Realizar un análisis detallado de las características del dataset, incluyendo estadísticas descriptivas, distribuciones y tendencias.
- **Análisis Exploratorio y Limpieza de Datos:**
  Realizar un completo análisis, que permita extraer insights e información relevante sobre las necesidades y el mercado actual.
- **Exploración de Correlaciones:**
  Investigar las relaciones entre variables clave, como experiencia, ubicación y nivel educativo, en relación con los salarios.
- **Identificación de Tendencias:**
  Descubrir posibles tendencias en el mercado laboral tecnológico en Argentina, como la demanda de habilidades específicas.
- **Modelos Predictivo:**
  Se podría desarrollar un modelo predictivo, utilizando modelos de regresión, para estimar salarios basándote en características específicas.

Objetivo Secundario

Si bien no es el cuore de este estudio, se ve con gran potencial, para agregar valor agregado al informe realizar el siguiente estudio:

- **Modelos de Clasificación:**
  Se podría clasificar a los profesionales según distinto tipo de categorías como ser: nivel educativo, crear perfiles laborales, tipo de empresas, ubicación geográfica, condiciones de pago, etc.

Adquisición de los Datos (Data Adquisition)

El dataset (DS) utilizado es de libre acceso y puede encontrarse en el siguiente link: Sysarmy Tecnología 2022. Este se encuentra en extensión ".csv" y consta de datos que surgen de una encuesta donde respondieron unas 5800 personas relacionadas al sector de interés. El mismo está compuesto por 5388 instancias y 44 variables.

Cabe aclarar que el mismo data de mediados del 2022.

Limpieza de los Datos (Data Wrangling)

El Data Wrangling implica la transformación y limpieza de datos en bruto para convertirlos en un formato adecuado para poder ser trabajado. Los datos en bruto suelen ser desordenados, inconsistentes e incompletos, lo que dificulta su análisis. Esta etapa es primordial y es, en conjunto con el EDA, donde se invertirá la mayor cantidad del tiempo (suele estimarse que equivale al 60% de dedicación dentro de un proyecto de este tipo).

Las etapas del proceso, en general, son: Importación de datos, Exploración de datos, Limpieza de datos, Estandarización, Combinación de datos, Transformación de datos, Validación de datos.

Información General de los Datos

A continuación, se muestra la información del dataframe crudo. Podemos ver el nombre de las columnas, la cantidad de no nulos y el tipo de dato:

#	Column	Non-Null Count	Dtype
0	work_country	5358 non-null	object
1	work_province	5358 non-null	object
2	work_dedication	5358 non-null	object
3	work_contract_type	5358 non-null	object
4	salary_monthly_BRUTO	5358 non-null	float64
5	salary_monthly_NETO	5358 non-null	float64
6	numero	5358 non-null	bool
7	salary_in_usd	1640 non-null	object
8	salary_last_dollar_value	1063 non-null	object
9	salary_pay_cripto	161 non-null	object
10	salary_%_cripto	789 non-null	object
11	salary_has_bonus	5358 non-null	object
12	salary_bonus_tied_to	5357 non-null	object
13	salary_inflation_adjustment	5358 non-null	object
14	salary_percentage_inflation_adjustment	5358 non-null	object
15	salary_month_last_inflation_adjustment	5358 non-null	object
16	salary_comparison_last_semester	5358 non-null	int64
17	salary_benefit	5358 non-null	object
18	salary_satisfaction	5357 non-null	float64
19	Trabajo de	5357 non-null	object
20	profile_years_experience	5357 non-null	float64
21	work_years_in_company	5357 non-null	float64
22	work_years_in_current_position	5357 non-null	float64
23	work_people_in_charge_of	5357 non-null	float64
24	tools_platform	5356 non-null	object
25	tools_programming_languages	5354 non-null	object
26	tools_frameworks	5354 non-null	object
27	tools_data_bases	5354 non-null	object
28	tools_qa_testing	5354 non-null	object
29	company_employee_number	5356 non-null	object
30	work_work_modality	5356 non-null	object
31	work_days_in_the_office	5356 non-null	float64
32	company_recommended	5356 non-null	float64
33	profile_studies_level	2659 non-null	object
34	profile_studies_level_state	2659 non-null	object
35	profile_career	2556 non-null	object
36	profile_university	2502 non-null	object
37	profile_boot_camp	693 non-null	object
38	profile_boot_camp_carrer	436 non-null	object
39	work_on_call_duty	1718 non-null	object
40	salary_on_call_duty_charge	1718 non-null	float64
41	work_on_call_duty_charge_type	1718 non-null	object
42	profile_age	5354 non-null	float64
43	profile_gender	5354 non-null	object

Figura 1: Información General del Dataset

Se realiza un 'describe' para tener una visión global del dataset por medio de algunos indicadores básicos, pero muy útiles para una primera reflexión:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
work_country	5358	1	Argentina	5358	NaN	NaN	NaN	NaN	NaN	NaN	NaN
work_province	5358	24	Ciudad Autónoma de Buenos Aires	2699	NaN	NaN	NaN	NaN	NaN	NaN	NaN
work_dedication	5358	2	Full-Time	5106	NaN	NaN	NaN	NaN	NaN	NaN	NaN
work_contract_type	5358	5	Staff (planta permanente)	4068	NaN	NaN	NaN	NaN	NaN	NaN	NaN
salary_monthly_BRUTO	5358.00	NaN	NaN	NaN	363510.54	543925.46	0.00	150000.00	256000.00	412657.00	28000000.00
salary_monthly_NETO	5358.00	NaN	NaN	NaN	277010.79	401942.30	0.00	118612.19	201000.00	294000.00	15000000.00
numero	5358	1	True	5358	NaN	NaN	NaN	NaN	NaN	NaN	NaN
salary_in_usd	1640	3	Cobro parte del salario en dólares	663	NaN	NaN	NaN	NaN	NaN	NaN	NaN
salary_last_dollar_value	1063	359	130	50	NaN	NaN	NaN	NaN	NaN	NaN	NaN
salary_pay_cripto	161	2	Cobro todo el salario criptomonedas	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN
salary_%_cripto	789	54	0	632	NaN	NaN	NaN	NaN	NaN	NaN	NaN
salary_has_bonus	5358	5	No	3022	NaN	NaN	NaN	NaN	NaN	NaN	NaN
salary_bonus_tied_to	5357	207	No recibo bono	2982	NaN	NaN	NaN	NaN	NaN	NaN	NaN
salary_inflation_adjustment	5358	5	Uno	1729	NaN	NaN	NaN	NaN	NaN	NaN	NaN
salary_percentage_inflation_adjustment	5358	342	0	1067	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figura 2: Primeros Indicadores mediante un 'describe'

Cabe aclarar que solo se muestran algunas instancias para hacer el informe más legible. Si necesitaran ver el resultado completo del script, pueden hacerlo entrando en el link de Google Colab aquí facilitado: https://colab.research.google.com/drive/

Outliers y Ceros

En esta primera aproximación hemos visto la cantidad de categorías de las variables con escalas nominal u ordinal (cualitativas), y su moda; y las medidas de posición y de variabilidad de los datos en las variables con escalas discretas o intervalares (cuantitativas).

En este nuevo análisis, la idea es analizar los datos para observar los casos perdidos y los ceros que existen en el dataset.

Conteo y cálculo porcentual de los nulos y ceros

salary_bonus_tied_to	1	salary_bonus_tied_to	0.02
salary_satisfaction	1	salary_satisfaction	0.02
Trabajo de	1	Trabajo de	0.02
profile_years_experience	1	profile_years_experience	0.02
work_years_in_company	1	work_years_in_company	0.02
work_years_in_current_position	1	work_years_in_current_position	0.02
work_people_in_charge_of	1	work_people_in_charge_of	0.02
company_recommended	2	company_recommended	0.04
work_days_in_the_office	2	work_days_in_the_office	0.04
work_work_modality	2	work_work_modality	0.04
tools_platform	2	tools_platform	0.04
company_employee_number	2	company_employee_number	0.04
tools_qa_testing	4	tools_qa_testing	0.07
profile_age	4	profile_age	0.07
profile_gender	4	profile_gender	0.07
tools_frameworks	4	tools_frameworks	0.07
tools_programming_languages	4	tools_programming_languages	0.07
tools_data_bases	4	tools_data_bases	0.07
profile_studies_level	2699	profile_studies_level	50.37
profile_studies_level_state	2699	profile_studies_level_state	50.37
profile_career	2802	profile_career	52.30
profile_university	2856	profile_university	53.30
work_on_call_duty	3640	work_on_call_duty	67.94
salary_on_call_duty_charge	3640	salary_on_call_duty_charge	67.94
work_on_call_duty_charge_type	3640	work_on_call_duty_charge_type	67.94
salary_in_usd	3718	salary_in_usd	69.39
salary_last_dollar_value	4295	salary_last_dollar_value	80.16
salary_%_cripto	4569	salary_%_cripto	85.27
profile_boot_camp	4665	profile_boot_camp	87.07
profile_boot_camp_carrer	4922	profile_boot_camp_carrer	91.86
salary_pay_cripto	5197	salary_pay_cripto	97.00

Figura 3: Valores Nulos|

salary_monthly_BRUTO	14	salary_monthly_BRUTO	0.26
salary_monthly_NETO	254	salary_monthly_NETO	4.74
profile_years_experience	530	company_recommended	6.98
work_years_in_company	1921	profile_years_experience	9.89
work_years_in_current_position	1944	salary_on_call_duty_charge	28.76
work_people_in_charge_of	3911	work_years_in_company	35.85
work_days_in_the_office	3644	work_years_in_current_position	36.28
company_recommended	374	work_days_in_the_office	68.01
salary_on_call_duty_charge	1541	work_people_in_charge_of	72.99

Figura 4: Ceros

Información de las Columnas

Con lo analizado hasta el momento, se esboza un panorama inicial de todas las columnas presentes en el dataframe:

0. **work_country/País del encuestado**: Argentina el 100% de la muestra (notar: no es país para el cual trabaja, sino que entendemos que es residencia).
1. **work_province/Provincia del encuestado**: Más de la mitad: CABA (idem nota anterior).
2. **work_dedication/Dedicación Laboral**: Una amplia mayoría trabaja Full-Time
3. **work_contract_type/Tipo de contrato**: Más de 5/6 de la muestra respondió "Staff Permanente". Habría 5 categorías diferentes.
4. **salary_monthly_BRUTO/Salario Bruto**: Variable con con alta variabilidad y extremos -> Analizar en detalle si se la quiere utilizar.
5. **salary_monthly_NETO/Salario Neto**, similar a Salario Bruto en cuanto a distribución, y además tiene 212 casos perdidos. -> Analizar para usar.
6. **numero**: Esta variable es del tipo booleano (True or False) y todas sus respuestas son "True". No se llega a entender su función -> Analizar si es necesario que integre el DS.
7. **salary_in_usd/Sueldo en Dólares**: Sólo respondió un 30% aprox. de la muestra total.
8. **salary_last_dollar_value/Último Salario en Dólares**: Sólo respondió un 20% aprox. de la muestra total.
9. **salary_pay_cripto**: Sólo respondíeron 161, muestra muy chica.
10. **salary_%_cripto**: Sólo respondíeron 789, muestra muy chica. De todas maneras, de estas, casi el 80% dijo no recibir ningún porcentaje en esta forma de pago.
11. **salary_has_bonus/Tiene bono el salario**: Más de la mitad de la muestra dijo "NO". Se tienen 5 valores únicos -> Revisar.
12. **salary_bonus_tied_to/Bono vinculado a**: Más de la mitad dijo "No recibo bono". Pero habría 207 categorías diferentes (revisar/analizar si se la quiere utilizar).
13. **salary_inflation_adjustment/Ajuste por inflación al salario**: Muestra 5 Categorias, y menos de la mitad dijo "NO".
14. **salary_percentage_inflation_adjustment/Porcentaje de ajuste por inflación al salario**: La variable tiene valores extremos superiores, analizar y limpiar si se la quiere utilizar.
15. **salary_month_last_inflation_adjustment/Último mes de ajuste por inflación**: El más frecuente es "Julio" con casi 1500 respuestas.
16. **salary_comparison_last_semester/Comparación salaro con último semestre**: respuesta en números que parece ser categórica -> investigar la definición de la variable o pregunta realizada en la encuesta en caso de querer seleccionarla para trabajar.
17. **salary_benefit/Beneficios Salariales**: Tiene 1886 valores únicos -> investigar la definición de la variable o pregunta realizada en la encuesta en caso de querer seleccionarla para trabajar.
18. **salary_satisfaction/Satisfacción**: Notar -> categórica ordinal (respuesta en números estilo escala).
19. **Trabajo de**: Tenemos 347 valores únicos (investigar) y un 45% trabaja como "Developer"
20. **profile_years_experience/Experiencia laboral en años**: la mitad de la muestra tiene como máximo 7 años de experiencia laboral. Se observan valores extremos superiores -> limpiar si se la quiere analizar.

21. **work_years_in_company/Años en la empresa**: la mitad de la muestra tiene como máximo 2 años en la empresa. Revisar valores máximos y limpiar.
22. **work_years_in_current_position/ Años en la posición actual**: la mitad de la muestra tiene como máximo 2 años en la empresa. Revisar valores máximos y limpiar.
23. **work_people_in_charge_of/Gente a cargo**: Esta variable registra una dispersión muy alta -> analizar en detalle/limpiar si se la quiere seleccionar para trabajar.
24. **tools_platform/Plataformas**: Habría 1197 plataformas distintas. La Moda fue "Ninguna de las anteriores" con 1013 respuestas. Es decir, hay mucha variabilidad y no es buen indicador que la moda sea "ninguna de las anteriores"-> Necesita trabajo de análisis detallado, depuración, agrupación de categorías si se la quiere utilizar para trabajar. Tiene sólo dos casos perdidos.
25. **tools_programming_languages/Lenguajes de programación** (1318 categ. y 885=Ninguno) -> Es una variable interesante, estudiar con mayor profundidad porque la cantidad de categ. parece un exceso
26. **tools_frameworks/framework** (1371 categ. y 1379=Ninguna)
27. **tools_data_bases/data bases** (1128 categ. y 1986=Ninguna)
28. **tools_qa_testing/qa testing** (539 categ. y 3099=Ninguna)
29. **company_employee_number/Número de empleados de la empresa**: Cargada como categórica con aparentemente 10 intervalos. Moda: Entre 11 y 50 empleados (con una proporción aproximada de 1/6 del total).
30. **work_work_modality/Modalidad de trabajo**: Más del 50% respondió que trabaja 100% remoto.
31. **work_days_in_the_office**: Los que asisten a la oficina, en su mayoría, va 1 vez a la semana
32. **company_recommended/Mejor empresa de la ciudad**: Registra todas las respuestas.
33. **profile_studies_level/Nivel de estudio**: La mayoría (>60% de la muestra) respondió Universitario, pero habría 7 categorías posibles y solo respondió la menos de la mitad del total.
34. **profile_studies_level_state/Estado del estudio**: Habría tres categorías, donde la moda para toda la muestra es "Completado", con un valor cercano al 50% de los datos.
35. **profile_career/Carrera**: Tiene más de la mitad de casos perdidos. Habría 391 carreras "diferentes" (revisar/analizar si se quiere utilizar). Moda: Ing.Sist.Inf. con aprox. 1/5 del total de respuestas no vacías.
36. **profile_university/Universidad**: Habría 450 universidades "diferentes" (revisar/analizar si se quiere utilizar). Moda: UTN con 586 casos.
37. **profile_boot_camp/Cursos**: Muestra muy chica respecto del total como para ser representativa (693 respuestas)
38. **profile_boot_camp_carrer/Cursos**: Muestra muy chica respecto del total como para ser representativa (436 respuestas)
39. **work_on_call_duty**: Tiene 3640 valores faltantes o nulos
40. **salary_on_call_duty_charge**:Tiene 3640 valores faltantes o nulos
41. **work_on_call_duty_charge_type**: Tiene 3640 valores faltantes o nulos

42. **profile_age/Edad**: La mitad de la muestra tiene como máximo 33 años. Se observan valores extremos superiores -> limpiar si se la quiere analizar.
43. **profile_gender/Género**: Amplia mayoría de hombres en la muestra. Arroja 14 valores únicos -> Revisar y limpiar para utilizar.

Tratamiento de Outliers y Ceros

En base a lo analizado se dispone a:

- Eliminar las variables con NaN>50%
- Volver a estudiar el DF y visualizar los NaN restantes para tomar decisiones al respecto
- Eliminar o modificar, en base a algún criterio a definir, las filas de NaN restantes

Observaciones del Data Wrangling

Tras estas modificaciones, el modelo preliminar queda de la siguiente manera:

#	Column	Non-Null Count	Dtype
0	work_country	5351 non-null	object
1	work_province	5351 non-null	object
2	work_dedication	5351 non-null	object
3	work_contract_type	5351 non-null	object
4	salary_monthly_BRUTO	5351 non-null	float64
5	salary_monthly_NETO	5351 non-null	float64
6	numero	5351 non-null	bool
7	salary_has_bonus	5351 non-null	object
8	salary_bonus_tied_to	5351 non-null	object
9	salary_inflation_adjustment	5351 non-null	object
10	salary_percentage_inflation_adjustment	5351 non-null	object
11	salary_month_last_inflation_adjustment	5351 non-null	object
12	salary_comparison_last_semester	5351 non-null	int64
13	salary_benefit	5351 non-null	object
14	salary_satisfaction	5351 non-null	float64
15	Trabajo de	5351 non-null	object
16	profile_years_experience	5351 non-null	float64
17	work_years_in_company	5351 non-null	float64
18	work_years_in_current_position	5351 non-null	float64
19	work_people_in_charge_of	5351 non-null	float64
20	tools_platform	5351 non-null	object
21	tools_programming_languages	5351 non-null	object
22	tools_frameworks	5351 non-null	object
23	tools_data_bases	5351 non-null	object
24	tools_qa_testing	5351 non-null	object
25	company_employee_number	5351 non-null	object
26	work_work_modality	5351 non-null	object
27	work_days_in_the_office	5351 non-null	float64
28	profile_age	5351 non-null	float64
29	profile_gender	5351 non-null	object

Figura 5: Dataset curado lego del DW

Análisis Exploratorio de los Datos (EDA)

Una vez que se tiene la base de datos limpia y curada, se realiza un segundo análisis de los datos. El principal fin de este análisis es comprender mejor los datos, identificar problemas que no hayan sido detectados en el DW y ayudar a responder las hipótesis planteadas y formular nuevas hipótesis. Además, esto permitirá seleccionar las técnicas de análisis de una manera más precisa.

Gráficos Univariados y Bivariados

A continuación, se realizarán distintos gráficos en busca de insights y responder algunas de las hipótesis:

Estudiando la Distribución de Edades

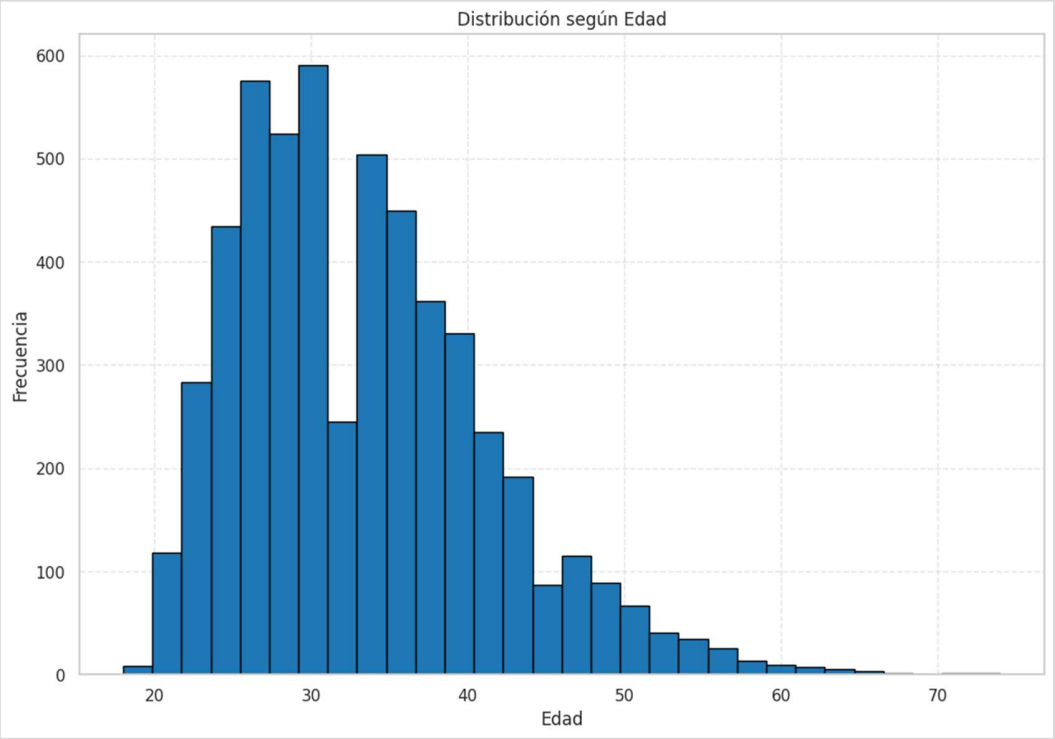

Figura 6: Distribución de las Edades en el Sector de IT

Estudiando la Experiencia Laboral en función del Sueldo Neto

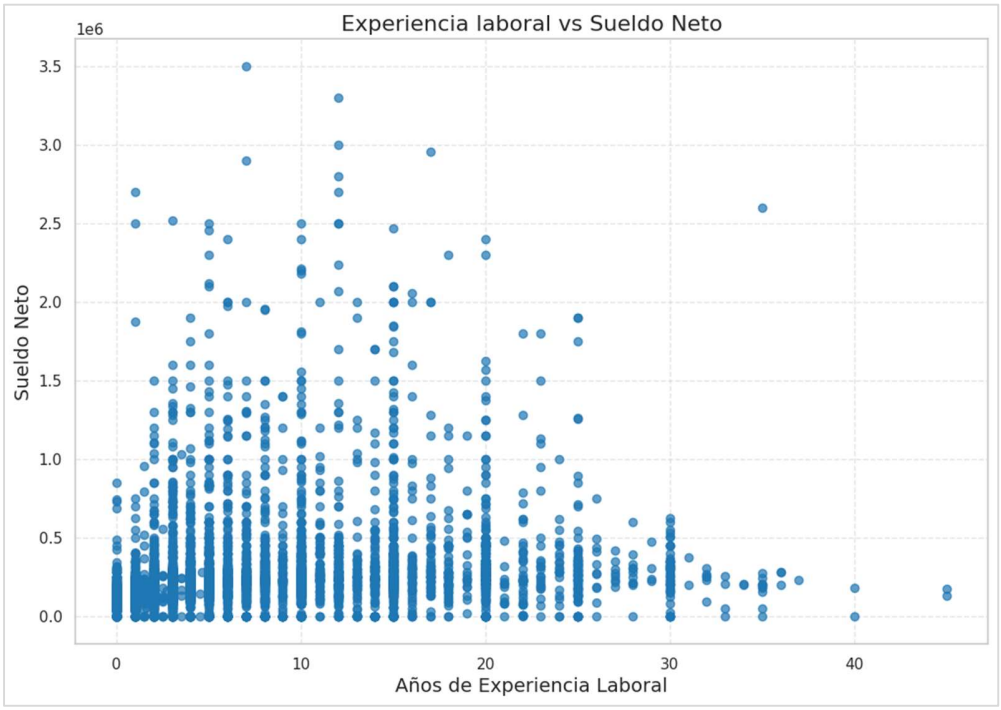

Figura 7: Sueldo Neto según la experiencia laboral

Estudiando la empleabilidad según Género en el Sector de IT

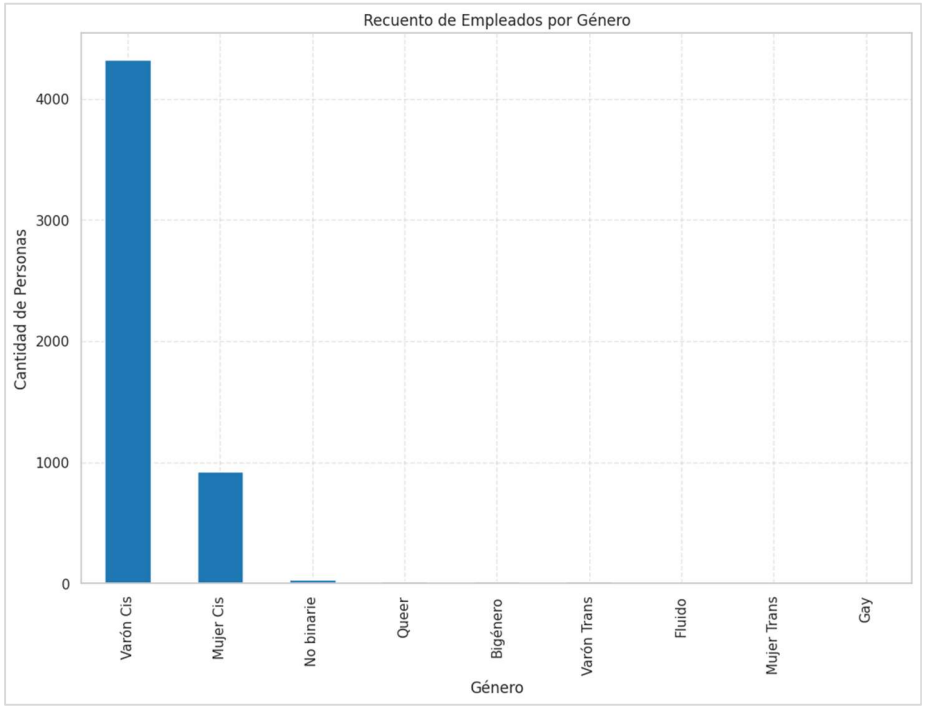

Figura 8: Distribución de los Géneros en el Sector de IT

Estudiando el Tipo de Contratación en el Sector IT

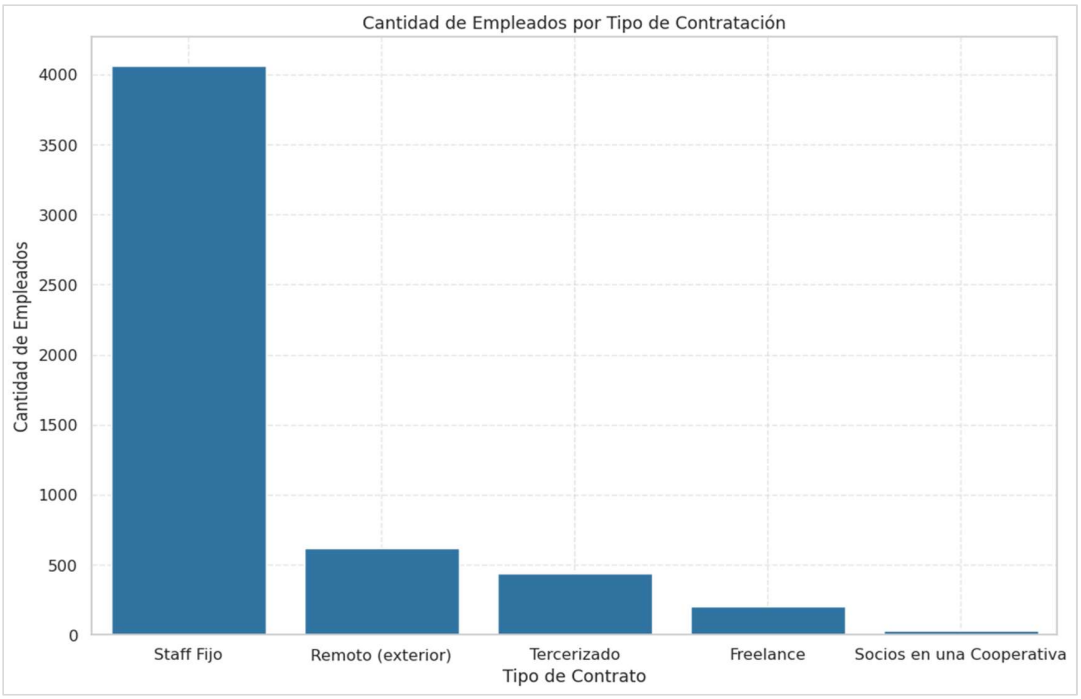

Figura 9: Tipo de Contratación más Frecuente

Estudiando el Distribución Geográfica de los Empleos

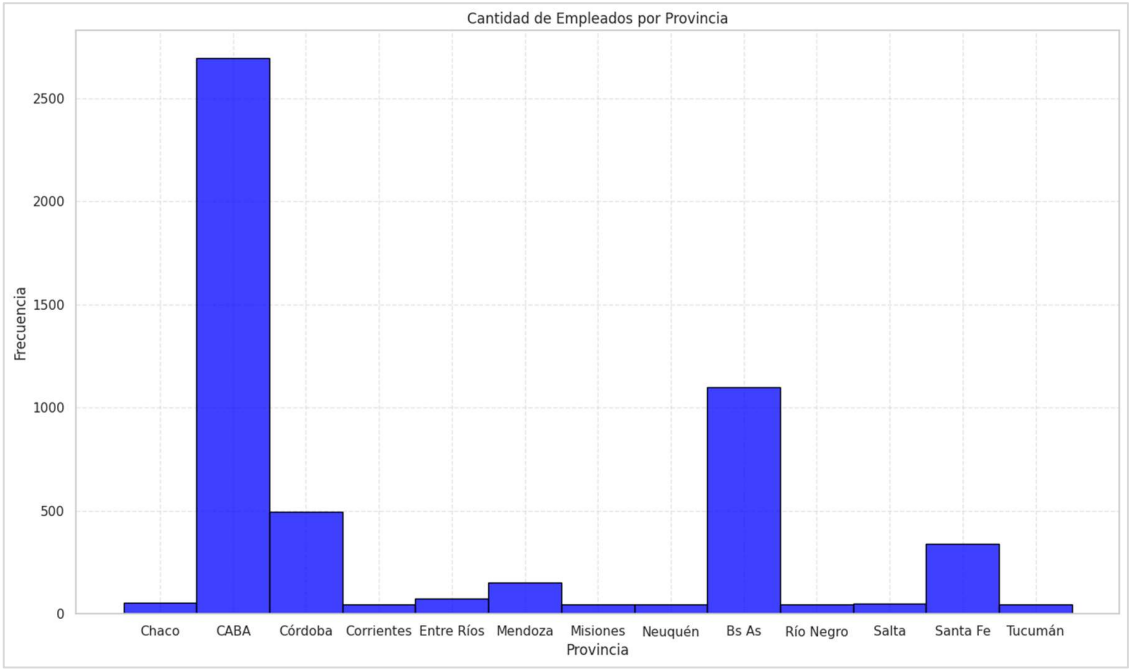

Figura 10: : Residencia de los Empleados del Sector de IT

Estudiando la Distribución del Salario Neto

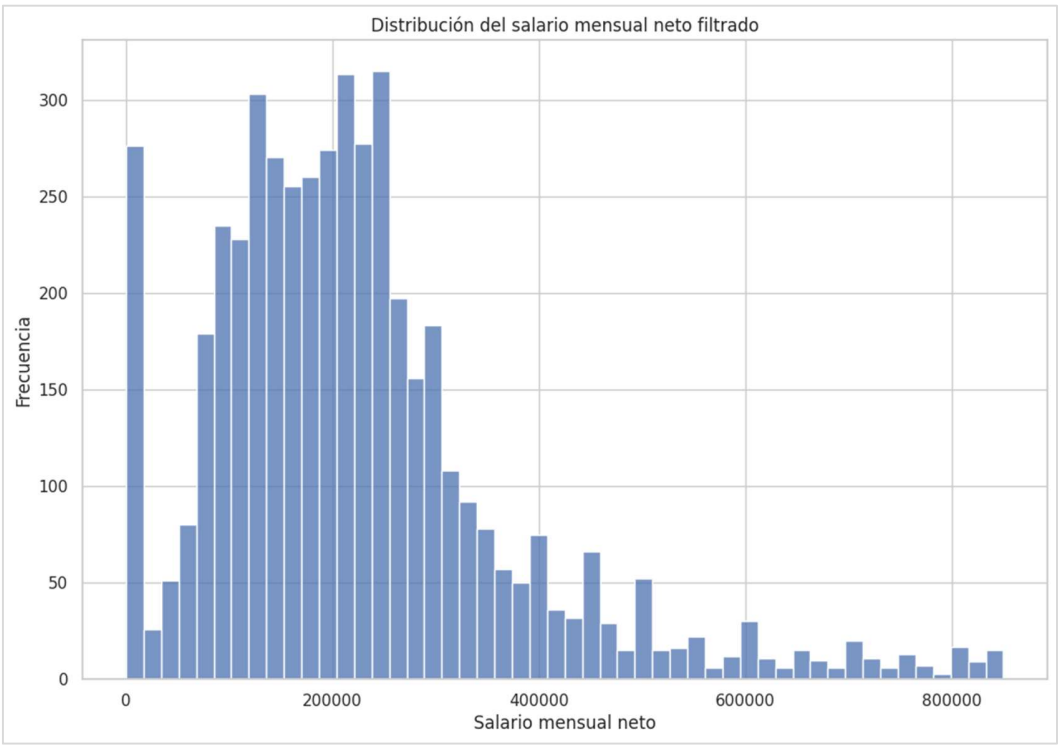

Figura 11: Distribución del Salario Neto

Análisis de la variable "lenguajes de programación"

Se trabajará en 2 aspectos fundamentales para nuestro estudio. Por un lado, la cantidad de ocurrencia de los lenguajes y por otro, el salario según lenguaje de programación

Estudiando la Distribución de Ocurrencias de Lenguajes

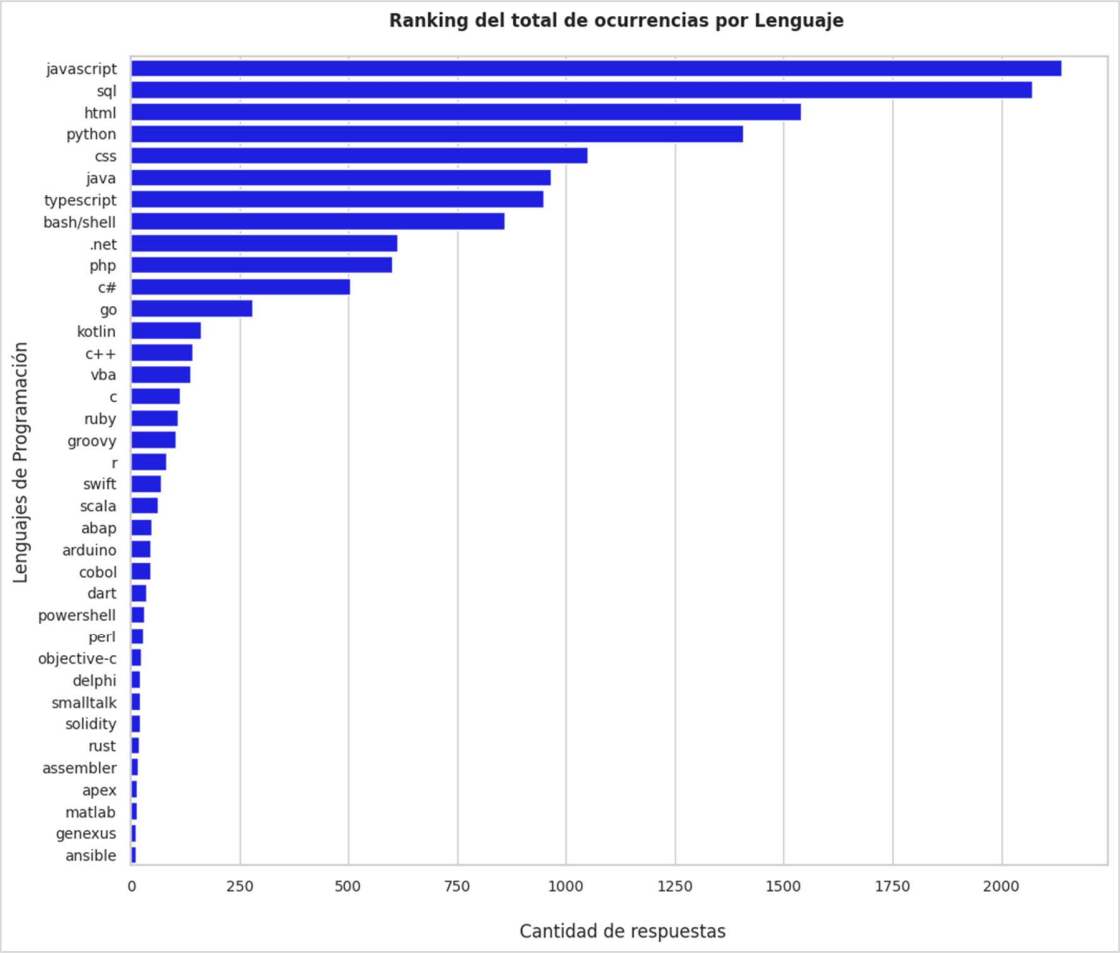

Figura 12: Distribución de la Ocurrencia según Lenguaje de Programación

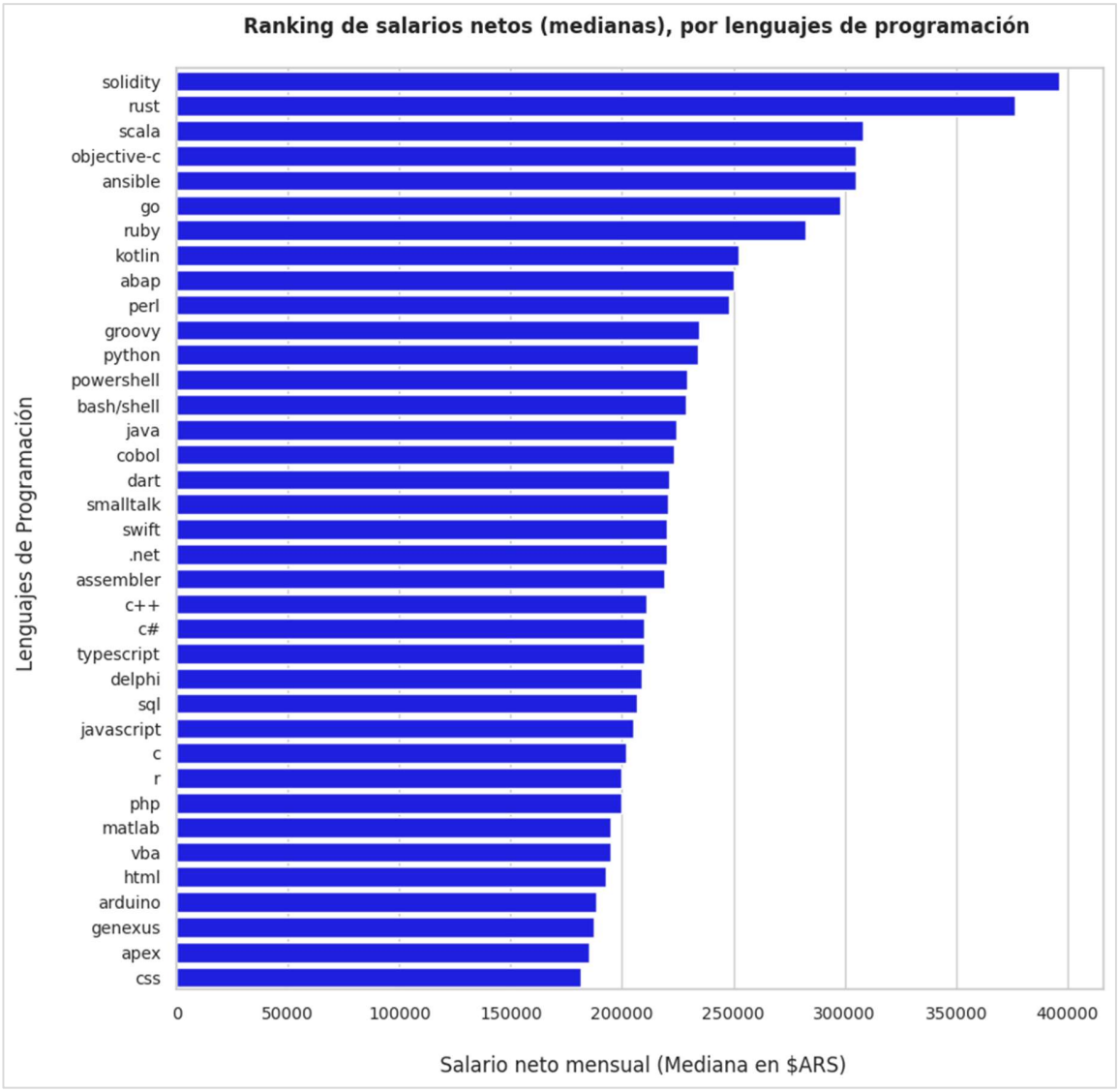

Figura 13: Salario según Lenguaje de Programación

Análisis de dos variables numéricas (Salario Bruto y Neto)

¿Por qué se estudia?

Se quiere decidir si sacar (o no) la columna de salario bruto, por lo que se busca si existe algún tipo de asociación entre ellas para limpiar aún más el dataframe y que sea lo más eficiente posible

¿Cómo se hará?

Se visualizará primero la distribución conjunta de ellas a través de un scatterplot.

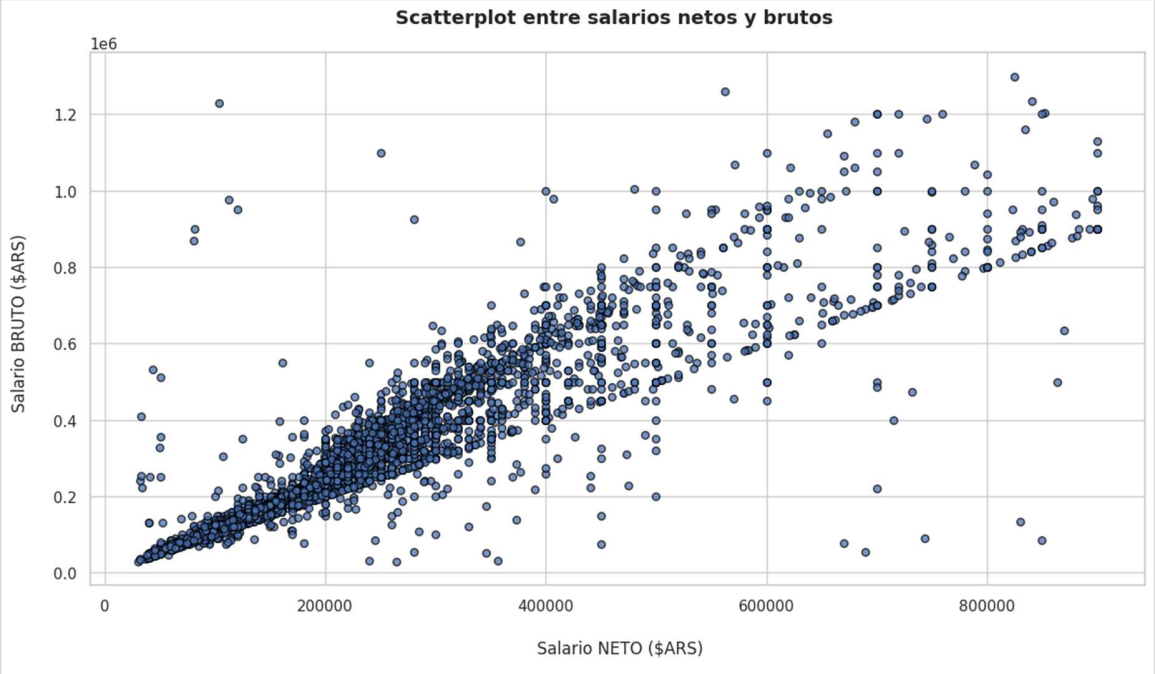

Figura 14: Salario Bruto vs Salario Neto

Estudio de algunos Indicadores

Se expondrán, a continuación, distintas correlaciones para que, en conjunto con el gráfico (Figura 14), ayudaran a entender mejor estas variables. Si se quiere ver la manera de cálculo ir a la Sección 7.2 del siguiente link: https://colab.research.google.com/drive/

- Correlación entre ambas variables es: 0.82
- Coef. de var. de Salario Neto: 1.19
- Coef. de var. de Salario Bruto: 1.48

Observaciones

- Al analizar los estudios realizados, parece convenir trabajar con una sola variable, ya que ambas están correlacionadas (ver coeficiente de correlación entre salario neto y bruto igual a 0,82).
- Se decide utilizar la variable 'salary_monthly_NETO' ya que tiene un coeficiente de variación menor al del salario bruto, lo que implica menor variación de los datos en torno a su media.
- Además, como dato útil, el salario neto muestra el dinero recibido en mano, el cual suele ser un dato más buscado que el salario bruto.

Conclusiones del EDA:

- En el histograma (Figura 6) se puede observar que la tendencia de edad de los empleados se encuentra en el rango de 25años<X<30años
- El gráfico, Figura 7, muestra como el salario va creciendo según los años de experiencia laboral, al menos en los 10-15 primeros años de experiencia. Por otro lado, se ve que los empleados con más de 15 años de experiencia suelen tener un sueldo base más alto.
- El género de Varón Cis es, por lejos, el que más representantes tiene entre los empleados en esta encuesta. Mientras que Mujer Cis lo sigue pero muy por debajo, casi 4 veces menos mujeres empleadas
- Se ve, claramente, en la Figura 9, que el tipo de contratación más habitual es el de Staff Fijo, mientras el resto de las opciones se encuentran 7 veces (o más) por debajo
- Las provincias con más empleo en el sector IT, según esta encuesta, son Buenos Aires, con CABA como principal contratador, seguidos por Córdoba y Santa Fe. Sería interesante correlacionar estas provincias con el sueldo que perciben. De esta manera, podremos tener una idea de la brecha salarial según el lugar geográfico dentro de la Argentina.
- Se puede observar, en la Figura 11, que el salario más usual ronda los 200.000ARS, mientras que tenemos una frecuencia grande en el 0, esto debe darse por gente que no completó la celda y se interpretó como cero la misma.
- Hay que notar que, tal como se observa en la Figura 12, "Solidity" tiene muy pocas observaciones, tal como ocurre con otros lenguajes; dificultando la posibilidad de considerar como representativo este resultado. Es difícil establecer una asociación entre la popularidad de los lenguajes de programación y los mejores salarios. Se necesitaría un muestreo más grande.
- El estudio de la Figura 14, y los indicadores, demuestra que existe una correlación marcada entre ambas variables y que la variable de Salario Mensual Neto es la mejor opción debido a su coef. de var. menor.
- El gráfico de dispersión entre el Salario Bruto y Neto que se muestra arriba (Figura 14) nos da la pauta de que existe una relación lineal positiva entre las variables, y que por lo tanto no son independientes.

Data Frame Versión Final

En la figura 15 se ve representado al dataframe actual:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5017 entries, 0 to 5016
Data columns (total 32 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   work_country                            5017 non-null   object
 1   work_province                           5017 non-null   object
 2   work_dedication                         5017 non-null   object
 3   work_contract_type                      5017 non-null   object
 4   salary_monthly_BRUTO                    5017 non-null   float64
 5   salary_monthly_NETO                     5017 non-null   float64
 6   numero                                  5017 non-null   bool
 7   salary_has_bonus                        5017 non-null   object
 8   salary_bonus_tied_to                    5017 non-null   object
 9   salary_inflation_adjustment             5017 non-null   object
 10  salary_percentage_inflation_adjustment  5017 non-null   object
 11  salary_month_last_inflation_adjustment  5017 non-null   object
 12  salary_comparison_last_semester         5017 non-null   int64
 13  salary_benefit                          5017 non-null   object
 14  salary_satisfaction                     5017 non-null   float64
 15  Trabajo de                              5017 non-null   object
 16  profile_years_experience                5017 non-null   float64
 17  work_years_in_company                   5017 non-null   float64
 18  work_years_in_current_position          5017 non-null   float64
 19  work_people_in_charge_of                5017 non-null   float64
 20  tools_platform                          5017 non-null   object
 21  tools_programming_languages             5017 non-null   object
 22  tools_frameworks                        5017 non-null   object
 23  tools_data_bases                        5017 non-null   object
 24  tools_qa_testing                        5017 non-null   object
 25  company_employee_number                 5017 non-null   object
 26  work_work_modality                      5017 non-null   object
 27  work_days_in_the_office                 5017 non-null   float64
 28  profile_age                             5017 non-null   float64
 29  profile_gender                          5017 non-null   object
 30  cured_salary_neto                       4533 non-null   float64
 31  cured_salary_bruto                      4623 non-null   float64
dtypes: bool(1), float64(11), int64(1), object(19)
memory usage: 1.2+ MB
```

Figura 15: Dataframe Actual

Se decide, luego de analizar todos los estudios realizados, eliminar las siguientes variables:

- **salary_monthly_BRUTO:** No aporta valor significativo
- numero: No se entiende la referencia
- **salary_has_bonus:** Falta de valores
- **salary_bonus_tied_to:** Falta de valores
- **salary_inflation_adjustment:** Falta de valores
- **salary_percentage_inflation_adjustment:** Falta de valores
- **salary_month_last_inflation_adjustment:** Falta de valores
- **salary_comparison_last_semester:** Falta de valores
- **salary_benefit:** Falta de valores
- **salary_satisfaction:** No aporta valor significativo
- **tools_platform:** No aporta valor significativo (muchos valores distintos)
- **tools_frameworks:** No aporta valor significativo (muchos valores distintos)
- **tools_data_bases:** No aporta valor significativo (muchos valores distintos)
- **tools_qa_testing:** No aporta valor significativo (muchos valores distintos)
- **tools_programming_languages:** No aporta valor significativo en los modelos de ML. Se ha estudiado la variable por separado en la sección 7.1 y se sacaron conclusiones fructíferas de su estudio
- **Trabaja de:** No aporta valor significativo en los modelos de ML. Se ha incluido en distintos modelos y nunca mejoró las métricas. (muchos valores distintos)
- **company_employee_number:** No aporta valor significativo, ya que se estudia el comportamiento de los empleados, no de las empresas
- **work_days_in_the_office:** No aporta valor significativo
- **work_country:** Su valor es siempre el mismo, Argentina
- **work_dedication:** No aporta valor significativo
- **cured_salary_neto:** Variable creada para un estudio puntual
- **cured_salary_bruto:** Variable creada para un estudio puntual

Esto deja al siguiente dataframe (figura 16) como la versión final:

```
RangeIndex: 5017 entries, 0 to 5016
Data columns (total 10 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   work_province                   5017 non-null   object
 1   work_contract_type              5017 non-null   object
 2   salary_monthly_NETO             5017 non-null   float64
 3   profile_years_experience        5017 non-null   float64
 4   work_years_in_company           5017 non-null   float64
 5   work_years_in_current_position  5017 non-null   float64
 6   work_people_in_charge_of        5017 non-null   float64
 7   work_work_modality              5017 non-null   object
 8   profile_age                     5017 non-null   float64
 9   profile_gender                  5017 non-null   object
dtypes: float64(6), object(4)
```

Figura 16: Dataframe Versión Final

Se puede observar que el dataframe con el que se va a trabajar en los modelos de ML tendrá una extensión de 5017 filas y 10 columnas. Estas se consideraron, luego de analizar a fondo el dataset incial, como las columnas más relevantes y con la mayor cantidad de datos certeros.

Modelos de Machine Learning

El estudio se divide en 2 grandes grupos. El principal estudio se basa en la predicción de sueldos futuros. El otro estudio que se realizó, si bien es un estudio secundario, es la clasificación según brecha salarial y tipo de contrato.

Regresión

Regresión Lineal

La regresión lineal es un método estadístico para modelar la relación entre una variable dependiente continua y una o más variables independientes. Se utiliza para predecir valores numéricos basados en un conjunto de datos. La fórmula básica de la regresión lineal es:

$$y=\beta 0+\beta 1x+\epsilon$$

donde y es la variable dependiente, x es la variable independiente, $\beta 0$ es la intersección (constante), $\beta 1$ es la pendiente del modelo, y ϵ es el término de error.

Luego de crear y entrenar el modelo, se obtuvieron las siguientes métricas:

- R2: 0.0099
- RMSE (sqrt): 137337.5
- RMSE(log): 11.8

Conclusión

Claramente, el modelo así planteado no estaría siendo útil para la predicción y análisis de ninguna de las variables en juego. Tiene un error muy grande y el R2 es muy bajo. Si bien lo bueno es saber que el modelo corre, deberá ser tratado para optimizarlo y poder utilizarlo.

Regresión Múltiple

La regresión múltiple es una extensión de la regresión lineal que permite predecir el valor de una variable dependiente utilizando múltiples variables independientes. La fórmula básica de la regresión múltiple es:

$$y=\beta 0+\beta 1x1+\beta 2x2+...+\beta nxn+\epsilon$$

donde y es la variable dependiente, $x1,x2,...,xn$ son las variables independientes, $\beta 0$ es la intersección, $\beta 1,\beta 2,...,\beta n$ son los coeficientes de las variables independientes, y ϵ es el término de error.

Luego de crear y entrenar el modelo, se obtuvieron las siguientes métricas:

- R2: 0.3044
- RMSE (sqrt): 192664.3

En la figura 17, se deja una muestra del poder de predicción del actual modelo.

	Actual	Predicted	Sesgo	Error_porc
3279	65000.00	145188.00	-80188.00	-123.37
3897	264000.00	218124.00	45876.00	17.38
3671	128730.00	217800.00	-89070.00	-69.19
3240	180000.00	50932.00	129068.00	71.70
1670	235000.00	206008.00	28992.00	12.34
...	...	...	...	...
3979	170200.00	94764.00	75436.00	44.32
2710	270000.00	325496.00	-55496.00	-20.55
1487	250000.00	252432.00	-2432.00	-0.97
88	185000.00	196592.00	-11592.00	-6.27
457	590000.00	529528.00	60472.00	10.25

Figura 17: Predicción por medio de Regresión Múltiple

Conclusión

Cómo se puede ver en las métricas y en la predicción, el modelo no es del todo confiable, pero mejora bastante respecto a la regresión lineal

Se ha probado con distintas variables, agregando y sacando, pero los resultados son muy similares. De hecho, agregando algunas variables el resultado empeora.

Clasificación

Clasificación según Tipo de Contrato (Random Forest)

El Random Forest es un algoritmo de aprendizaje supervisado que consiste en un conjunto de árboles de decisión entrenados con diferentes subconjuntos del conjunto de datos. El modelo toma la predicción de cada árbol y la combina para obtener una predicción final más robusta y precisa. En clasificación, la predicción final se toma por mayoría de votos de los árboles individuales. Random Forest es conocido por su capacidad de manejar grandes cantidades de datos, detectar relaciones no lineales y reducir el sobreajuste.

Luego de crear y entrenar el modelo, se obtuvieron las siguientes métricas:

	precision	recall	f1-score	support
1	0.82	0.95	0.88	790
2	0.64	0.42	0.51	91
3	0.25	0.04	0.08	89
4	0.25	0.06	0.10	31
5	0.00	0.00	0.00	3
accuracy			0.79	1004
macro avg	0.39	0.30	0.31	1004
weighted avg	0.73	0.79	0.75	1004

Figura 18: Métricas para modelo de clasificación por Random Forest

Para simplificar la lectura de los resultados obtenidos en base a métricas, se aclara que los números corresponden a lo siguiente:

- 1: Staff Fijo
- 2: Remoto
- 3: Tercerizado
- 4: Freelance
- 5: Socios en una Cooperativa

Conclusión

Al parecer, se puede predecir mejor algunos tipos de contratación (staff fijo, Remoto) que otros. De todas maneras, el modelo es bastante confiable, ya que tiene un accuracy del 80%.

No se puede asegurar, pero pareciera que el problema está directamente relacionado con la cantidad de datos que se tienen da cada opción.

Clasificación según Brecha Salarial (Decision Tree)

Los árboles de decisión son un método de aprendizaje supervisado utilizado para problemas de clasificación y regresión. En clasificación, un árbol de decisión divide el conjunto de datos en subconjuntos más pequeños basados en la característica que proporciona la mayor ganancia de información en cada paso, hasta llegar a una predicción final. Cada nodo del árbol representa una característica, cada rama representa una decisión, y cada hoja representa un resultado o clase.

Los árboles de decisión son fáciles de interpretar y visualizar, pero pueden ser propensos al sobreajuste si no se podan adecuadamente.

Luego de crear y entrenar el modelo, se obtuvieron las siguientes métricas:

	precision	recall	f1-score	support
alta	0.43	0.31	0.36	105
baja	0.36	0.31	0.33	207
media	0.74	0.80	0.77	692
accuracy			0.65	1004
macro avg	0.51	0.47	0.49	1004
weighted avg	0.63	0.65	0.64	1004

Figura 19: Métricas para modelo de clasificación por Decision Tree

Para simplificar la lectura de los resultados obtenidos en base a métricas, se aclara a que valor numérico de brecha salarial corresponde cada categoría:

- alta: 500.000 < Sueldo
- media: 120.000 < Sueldo < 500.000
- baja: 0 < Sueldo < 120.000

Conclusión

El modelo da un accuracy de 63%, lo cual no estaría dentro del rango de confianza esperable para poder utilizarlo. Se ha probado con distintas variables, pero el resultado no varía manera favorable.

Por otro lado, se trabajó variando el rango de la brecha salarial, y ahí si se encuentran cambios. De todos modos, se tomará el rango actual porque es el más equilibrado de acuerdo a los datos sobre los sueldos de mediados de 2022, fecha de cuando data este dataset.

Luego de varias pruebas, se concluye que el problema radica en la cantidad de datos porque, al cambiar los rangos de las brechas, se ven cambios. Por ej.: Al hacer más grande el rango medio, el accuracy llega a un 80%, el tema es que predice con esa precisión para el rango medio, mientras que el rango bajo y alto, quedan en valores despreciables.

Conceptos Básicos sobre las Métricas Utilizadas

En el ámbito del aprendizaje automático, la evaluación del rendimiento de un modelo es crucial para determinar su eficacia y confiabilidad. Las métricas de evaluación proporcionan información valiosa sobre la capacidad del modelo para generalizar a nuevos datos y cumplir con los objetivos del problema en cuestión. A continuación, se ofrece una descripción de algunas métricas de evaluación comúnmente utilizadas, enfocándose en su interpretación y aplicación en modelos de clasificación y regresión.

- **Accuracy (exactitud):**

Mide el porcentaje de casos que el modelo ha acertado. Es la medida más directa de la calidad de los clasificadores, los valores se encuentran entre 0 y 1, y mientras más cerca de 1 mejor. Es una métrica en la cual no se debe confiar de manera plena, ya que puede ocasionar problemas si las clases de variables de destino en los datos no están balanceadas.

- **Precision (Precisión):**

Sirve para medir la calidad del modelo de machine learning en tareas de clasificación. Identifica qué porcentaje de valores que se han clasificado como positivos lo son realmente.

- **Recall (Exhaustividad):**

Esta métrica informa sobre la cantidad que el modelo de machina learning es capaz de identificar.

- **F1 Score (Puntaje F1):**

El puntaje F1 combina la precision y el recall en una sola medida, proporcionando un equilibrio entre ambas métricas. Es particularmente útil cuando ambas métricas son importantes, pero no necesariamente en la misma medida. Un modelo con alto F1 score indica un buen desempeño en ambos aspectos, identificando correctamente positivos y evitando falsos positivos.

- **R2 (Coeficiente de Determinación):**

El coeficiente de determinación (R2) es una métrica específica para modelos de regresión que mide la fuerza de la relación lineal entre las variables predichas y reales. Un valor de R2 cercano a 1 indica que el modelo explica una alta proporción de la variabilidad de los datos. Sin embargo, es importante tener en cuenta que un alto R2 no garantiza un buen modelo, ya que no considera la distribución del error ni la magnitud de los errores individuales.

- **RMSE (Error Cuadrático Medio):**

El error cuadrático medio (RMSE) es una métrica de error absoluto que mide la magnitud promedio del error entre las variables predichas y reales. Es frecuente utilizarla en modelos de Regresión. Un RMSE bajo indica que el modelo predice valores cercanos a los valores reales en promedio. Sin embargo, el RMSE es sensible a valores atípicos, por lo que se debe considerar en conjunto con otras métricas.

Modelos de Ensamble/Bostting y Mejoras

Los modelos de ensemble son técnicas de aprendizaje automático que combinan las predicciones de múltiples modelos base para mejorar el rendimiento general. La idea detrás de los métodos de ensemble es que, al combinar varios modelos, se puede reducir la varianza, el sesgo y mejorar la capacidad de generalización del modelo. Los dos enfoques principales en los modelos de ensemble son bagging y boosting. Este estudio se centrará en los modelos del tipo boosting.

XGBoost con Grid Search

XGboost Implementa modelos de predicciones débiles, con el objetivo de que secuencialmente cada modelo débil le permita ir ganando más poder predictivo hasta llegar a un modelo más robusto con mayor estabilidad en sus resultados.

Por su parte, Random Grid Search, es el método más común de ajuste de hiperparámetros. Este crea una cuadrícula de ajuste con combinaciones únicas de valores de hiperparámetros y utiliza Cross Validation para evaluar su rendimiento.

Luego de haber creado y entrenado el modelo, los resultados fueron:

XGboost

- Mean Squared Error: 35.038.526.615
- R2 Score: 0.35

XGboost con Grid Search

- Mean Squared Error: 34.267.976.442
- R2 Score: 0.37

Conclusiones Parciales

Basado en los resultados se puede concluir lo siguiente:

Regresión

El modelo de regresión múltiple arroja mejores resultados que el modelo de regresión simple. Se debería seguir trabajando sobre el mismo porque, si bien el resultado es mejor, aún no llega a lograr una robustez deseada para poder implementarlo a niveles usuario o producción.

Clasificación

Si bien cada modelo se evalúa con un objetivo diferente, uno está enfocado a la clasificación según tipo de contrato y el otro según brecha salarial, ambos arrojan unos primeros resultados alentadores. Podrían implementarse.

Modelos de Ensamble y Mejoras

Los resultados al aplicar XGboost son prácticamente los mismos que los de la regresión múltiple aplicada para la predicción de salarios del sector IT.

Por otro lado, si bien hay una mínima mejora en los resultados al aplicar Grid Search, los resultados siguen sin ser alentadores. Por otro lado, al aplicar Grid Search la velocidad de ejecución es bastante más lenta. Este factor hace que, a modo de conclusión sobre la aplicación de estos modelos, el modelo XGboost sin Grid Search sea el elegido por sobre el mismo con Grid Search.

Conclusiones Finales

- Las etapas de Data Wrangling y EDA se consideran muy satisfactorias y pueden verse los estudios al final de cada etapa (sección 6 y 7), donde se enumeran los resultados y conclusiones parciales de cada análisis.

- En los modelos de regresión, el que mejor resultados arrojó fue el de RegresiónMúltiple.

- En los modelos de clasificación, ambos modelos se comportaron de manera similar, y podrían ser utilizados.

- Las métricas no arrojan resultados óptimos en ninguno de los modelos.

- Se han aplicado modelos de ensemble como XGboost y mejoras como grid search, sin mejoras significativas para el estudio.

- A modo general, la aplicación efectiva de la ciencia de datos en el mercado laboral del sector de IT puede contribuir a la paridad salarial, la optimización de la estrategia de compensación y la toma de decisiones más informadas.

- Luego de diversos estudios, se concluye que el principal factor del bajo rendimiento del modelo, radica en la escasa cantidad de datos.

Lineas Futuras

- **Crear Variables Sintéticas:**
  Generar nuevas características a partir de las existentes, como interacciones entre variables o transformaciones no lineales.

- **Incorporar Más Datos:**
  Se cree conveniente poder vincular los datos de las encuestas Sysarmy de otros semestres y años, para poder tener un dataset más robusto.

- **Clusterización:**
  Es posible utilizar modelos de clusterización para lograr un mejor entendimiento de los grupos demográficos y optimizar los recursos. Esta podría ser una herramienta fundamental para abordar el problema de la brecha salarial y desarrollar políticas inclusivas.