



UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES

MÉTODOS QUANTITATIVOS PARA ANÁLISE MULTIVARIADA

DOCENTE: MARCELO LAURETTO

ATIVIDADE 1

RELATÓRIO DA APLICAÇÃO DE MODELOS DE REGRESSÃO

Gustavo Pompermayer Fulanetti Silva

NºUSP 14760280

João Gabriel de Senna Lamolha

NºUSP 14777879

SÃO PAULO 2024

1. Introdução

Este relatório apresenta a Atividade 1 da disciplina ACH2036 - Métodos Quantitativos para Análise Multivariada (MQAM), que constitui a aplicação de um modelo de Regressão Linear Múltipla. A base de dados utilizada, "2020_Data_Professional_Salary_MultipleFeatures", foi obtida da fonte <https://rafaeldontalgoncalez.com/2021/02/regressao-linear-multipla-em-python>

O objetivo central do estudo é avaliar a associação entre o salário anual em dólares (SalaryUSD), nossa variável resposta contínua, e um conjunto de variáveis preditoras que descrevem o perfil profissional, técnico e demográfico dos respondentes. O modelo buscará identificar e quantificar como fatores como YearsWithThisDatabase (anos de experiência), ManageStaff (gestão de equipe) e Country (País) influenciam a remuneração.

2. Detalhes da Base

A base de dados é composta por 8.627 registros de profissionais de dados, sem nenhum valor ausente. As variáveis utilizadas para a construção do modelo são detalhadas a seguir.

Variável Resposta (Dependente):

A. SalaryUSD - *Salário anual em dólares (USD)*

- Tipo: Numérica (contínua)
- Descrição: Valor do salário anual bruto do profissional, convertido para dólares americanos (USD).

Variáveis Preditoras (Independentes):

B. Country - *País*

- Tipo: Categórica
- Descrição: Indica o país de residência do profissional de dados participante da pesquisa.

C. PrimaryDatabase - *Banco de dados principal*

- Tipo: Categórica
- Descrição: Informa qual é o banco de dados mais utilizado ou preferido pelo profissional em seu trabalho (por exemplo: SQL Server, Oracle, MySQL, PostgreSQL, etc.).

D. YearsWithThisDatabase - *Anos de experiência com este banco de dados*

- Tipo: Numérica (inteira)
- Descrição: Representa o número de anos que o profissional trabalha com o banco de dados indicado em *PrimaryDatabase*.

E. EmploymentStatus - *Situação de emprego*

- Tipo: Categórica
- Descrição: Indica a condição atual de emprego do respondente (por exemplo: empregado em tempo integral, meio período, autônomo, desempregado, estudante, etc.).

F. ManageStaff - *Gerencia equipe*

- Tipo: Categórica (binária)
- Descrição: Indica se o profissional ocupa um cargo de gestão de pessoas (por exemplo: “Sim” se supervisiona outros funcionários, “Não” caso contrário).

G. DatabaseServers - *Servidores de banco de dados utilizados*

- Tipo: Categórica (múltipla escolha ou texto)
- Descrição: Lista ou identifica quais servidores de banco de dados o profissional utiliza em seu ambiente de trabalho.

H. PopulationOfLargestCityWithin20Miles - *População da maior cidade num raio de 20 milhas*

- Tipo: Numérica (contínua)
- Descrição: Representa o tamanho populacional da maior cidade localizada a até 20 milhas da residência do respondente, indicando o nível de urbanização da região.

I. Gender - *Gênero*

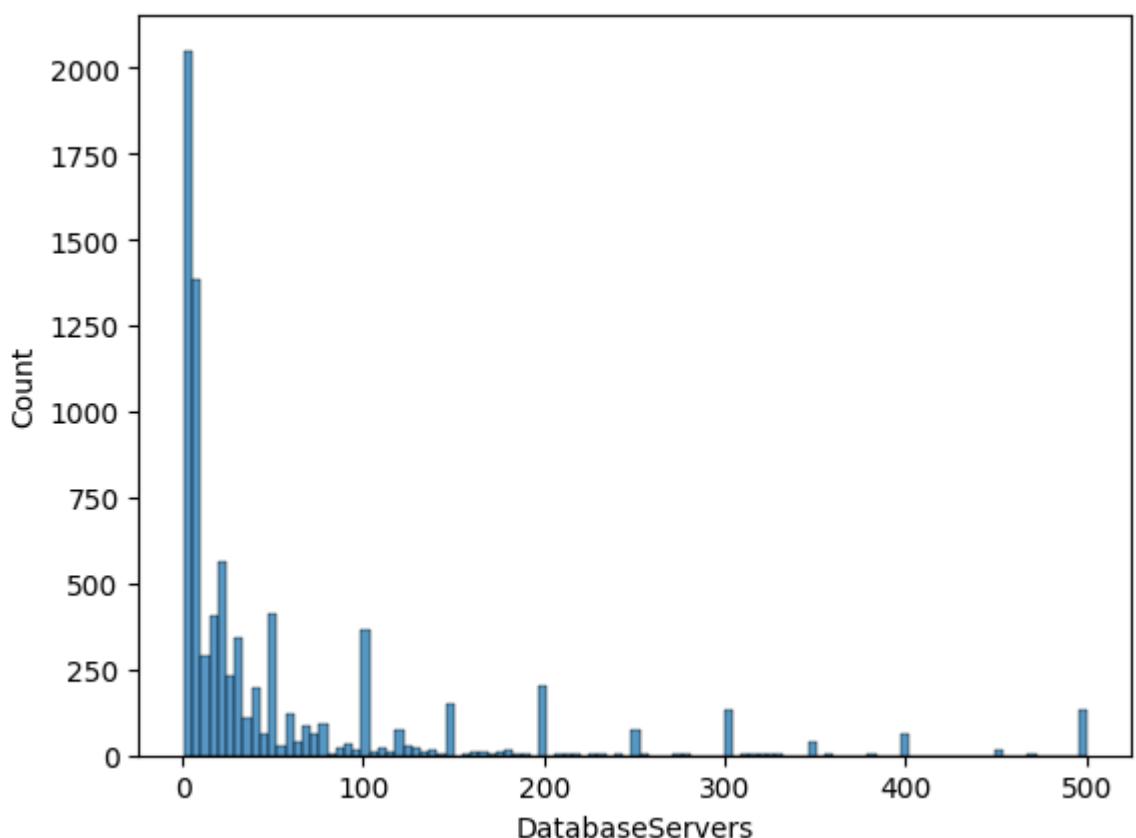
- Tipo: Categórica
- Descrição: Indica o gênero com o qual o profissional se identifica (por exemplo: masculino, feminino, outro, prefiro não dizer).
- Observação: Embora presente na base de dados original, esta variável foi proativamente excluída da análise. A decisão foi tomada para evitar a introdução de um potencial viés de gênero no modelo de remuneração, garantindo a ética do modelo.

3. Análises exploratórias

a. Assimetrias nas distribuições das variáveis numéricas:

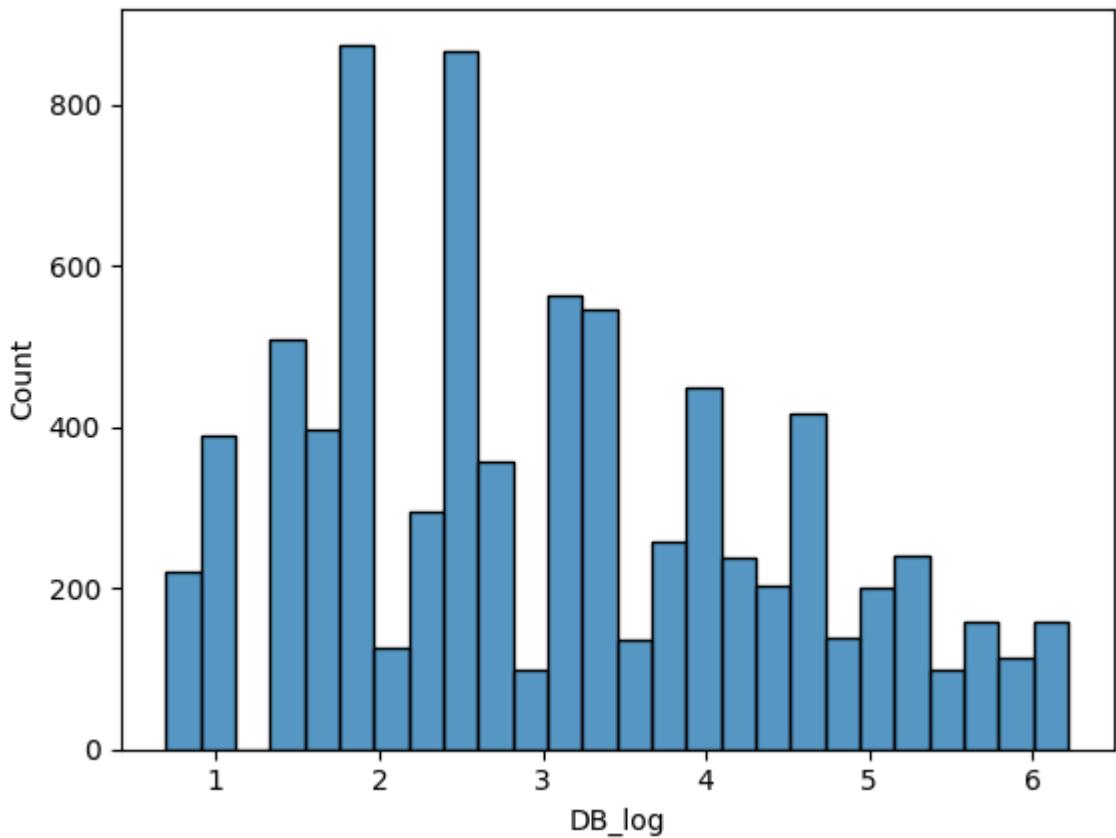
Para garantir a robustez e a validade do nosso modelo de regressão, foi realizada uma análise da distribuição das três principais variáveis numéricas: *DatabaseServers*, *YearsWithThisDatabase* e *SalaryUSD*. O objetivo era identificar e corrigir assimetrias severas, que podem violar os pressupostos do modelo. A análise baseou-se nas métricas de assimetria (skewness) e curtose (kurtosis), bem como na inspeção visual dos histogramas.

1. Servidores de Banco de Dados (*DatabaseServers*)



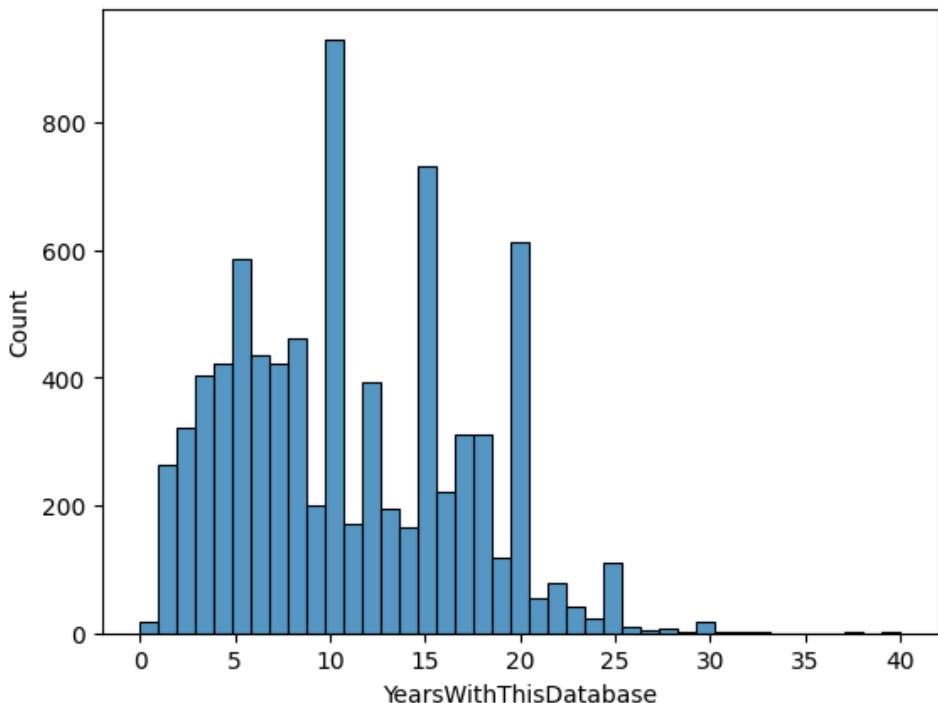
A variável original *DatabaseServers* apresentava uma assimetria positiva severa ($\text{Skew} = 2.84$). Isso indica que a grande maioria dos dados estava concentrada em valores baixos (poucos servidores), com uma longa cauda de valores extremos (muitos servidores).

- Ação: Foi aplicada uma transformação logarítmica, resultando na variável *DB_log*.



- Resultado: A transformação foi altamente eficaz. A assimetria foi neutralizada, caindo para 0.41, um valor considerado ideal (próximo de zero). O histograma de *DB_log* confirma isso, mostrando uma distribuição muito mais equilibrada e simétrica, adequada para a modelagem.

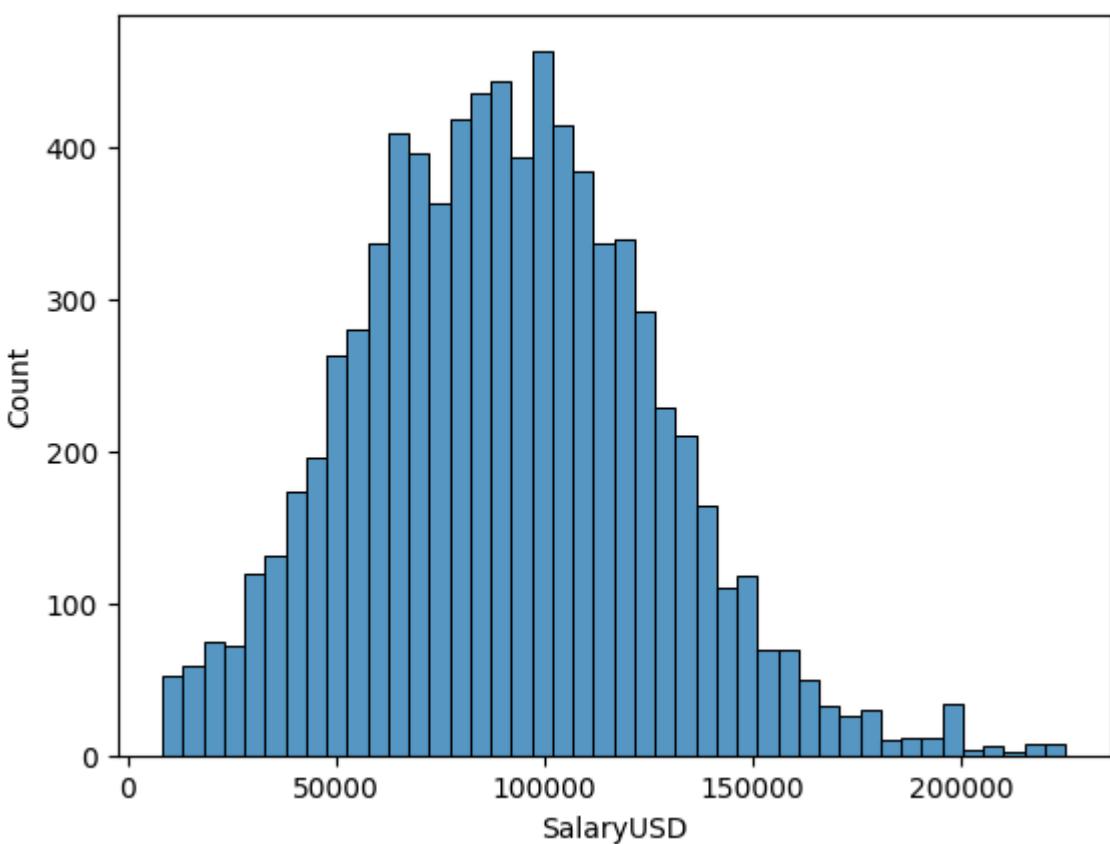
2. Anos de Experiência (*YearsWithThisDatabase*)



Diferente das outras, a variável *YearsWithThisDatabase* já se mostrou adequada em sua forma original.

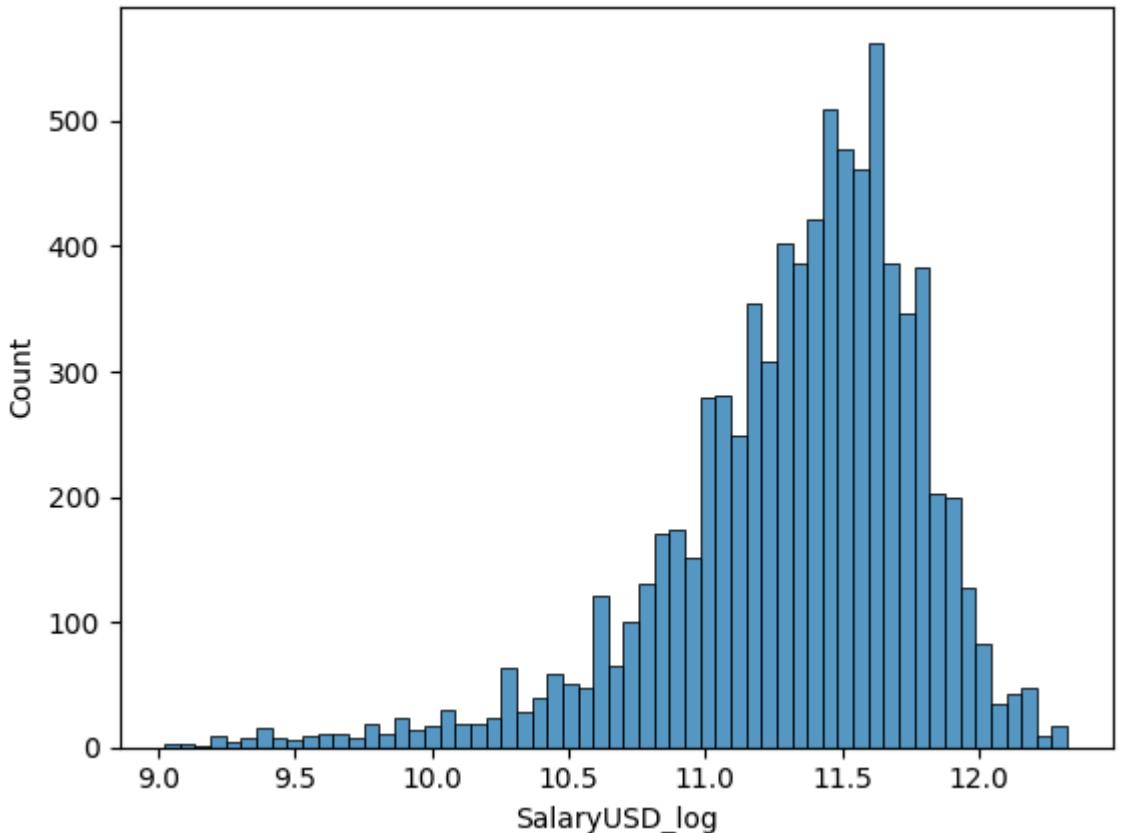
- Ação: Nenhuma transformação foi necessária.
- Resultado: A variável possui uma assimetria muito baixa ($\text{Skew} = 0.38$), indicando que sua distribuição já é satisfatoriamente simétrica. O histograma revela uma natureza multimodal (com vários picos, notavelmente em 10 e 15 anos), o que é uma característica natural dos dados (possivelmente refletindo diferentes "gerações" de profissionais) e não um problema estatístico que precise de correção.

3. Salário (*SalaryUSD*)



A análise da variável-alvo, o salário, revelou a descoberta mais importante.

Salary USD (Original, pós-clipping): A variável, após a remoção de outliers, apresentou uma distribuição quase perfeitamente normal ($\text{Skew} = 0.31$). Este é o cenário ideal para uma variável resposta em regressão linear.



SalaryUSD_log (Pós-transformação): Por curiosidade, tentamos aplicar uma transformação logarítmica. Como mostram o histograma e as métricas, essa ação foi contraproducente. A transformação introduziu uma nova assimetria negativa significativa ($\text{Skew} = -1.24$), distorcendo a distribuição para a esquerda. No entanto, ao aplicar a regressão linear, os resultados com a variável transformada foram consideravelmente melhores.

Conclusão da Análise:

Para o modelo final, foram utilizadas as variáveis *DB_log* e *YearsWithThisDatabase* como preditores, pois ambas possuem distribuições simétricas. Como variável resposta (alvo), foi utilizada a *SalaryUSD_log* (pós-clipping), pois tanto o RMSE e o R^2 se saíram melhor.

Para essas duas transformações log, usamos o `log1p` do numpy.

b. Necessidade de agrupamento para variáveis categóricas

Country (País)

A variável *Country* apresentava alta cardinalidade, com 93 países distintos. Manter todas essas categorias resultaria em um modelo com excesso de variáveis (colunas dummy), o que poderia prejudicar a performance e a generalização.

- Solução: Adotamos uma estratégia de agrupamento por frequência. Foi definido um limiar de 30 observações; qualquer país que apareceu menos de 30 vezes no dataset foi consolidado em uma única categoria geral, "Outros". Isso reduziu drasticamente a dimensionalidade, mantendo a relevância das localizações mais frequentes.

PopulationOfLargestCityWithin20Miles

Esta variável, embora categórica, possui uma natureza inherentemente ordenada (ex: "Menos de 100k", "100k-500k", "Mais de 1M").

Solução: Aplicamos uma Codificação Ordinal (Ordinal Encoding). A decisão se baseou na hipótese de que centros urbanos com maior população estão associados a um maior custo de vida e, consequentemente, a salários mais elevados. As categorias foram mapeadas para valores numéricos crescentes (ex: 0, 1, 2...), permitindo que o modelo capture essa relação de "peso" populacional.

EmploymentStatus

A variável *EmploymentStatus* continha múltiplas entradas de texto que, embora escritas de forma diferente, tinham o mesmo significado semântico.

Solução: Realizamos uma consolidação de categorias para normalizar as respostas. Por exemplo, entradas como 'Independent consultant, contractor, freelancer, or company owner' e 'Independent or freelancer or company owner' foram ambas mapeadas para a categoria unificada 'Independent/Freelancer'. Isso fortalece o poder preditivo da categoria ao evitar a fragmentação desnecessária.

PrimaryDatabase:

Assim como *Country*, esta variável também sofria de alta cardinalidade, com muitas tecnologias de banco de dados aparecendo pouquíssimas vezes.

- Solução: Foi aplicada a mesma técnica de agrupamento por frequência. Bancos de dados com baixa representatividade (poucas aparições) foram agrupados na categoria genérica 'DB_OtherDB'.

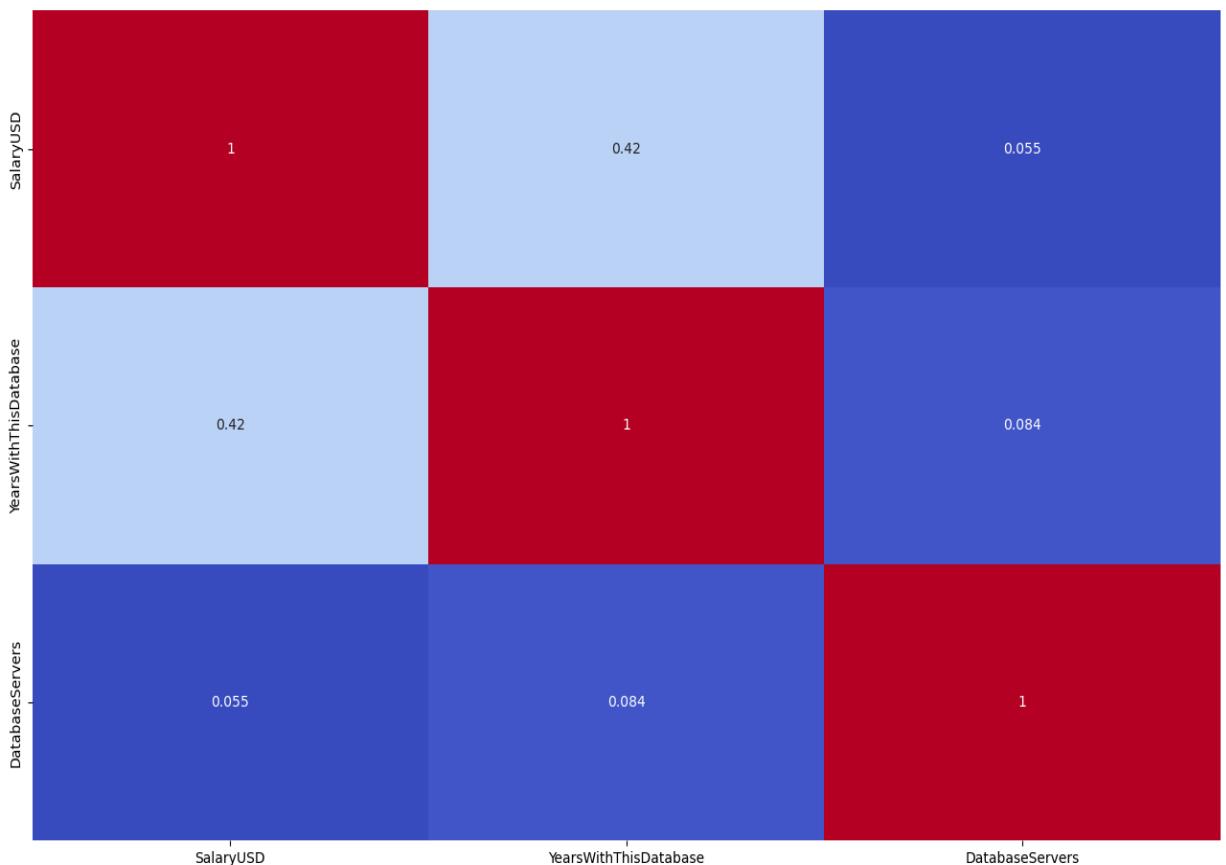
YearsWithThisDatabase:

Esta é uma variável numérica, mas sua relação com o salário provavelmente não é linear.

- Solução: Optamos por discretizar a variável, ou seja, convertê-la em faixas categóricas (ex: "0-5 anos", "6-10 anos"). A lógica é que o impacto no salário ao passar de 1 para 2 anos de experiência é muito mais significativo do que ao passar de 20 para 21 anos. O agrupamento em faixas permite que o modelo linear capture melhor essa relação não-linear, tratando cada faixa de experiência com um peso diferente.

c. Associações entre as principais variáveis

Antes da transformação das variáveis em faixas, analisamos a matriz de correlação (heatmap) entre as principais variáveis numéricas do modelo (já com o clipping de outliers aplicado). O objetivo foi identificar o grau de associação linear entre elas.

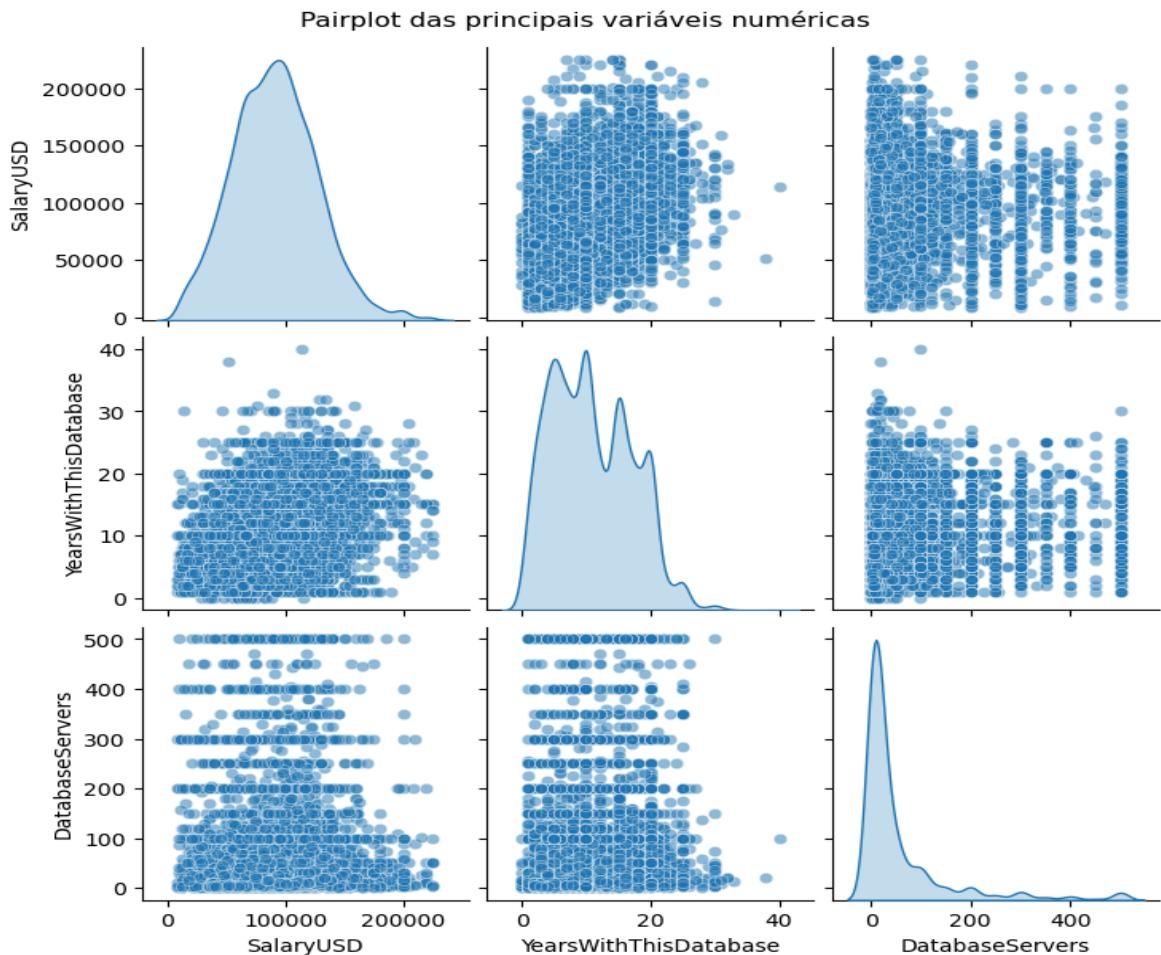


A análise do heatmap revela:

- Correlação Moderada (Salário vs. Experiência): A correlação entre *SalaryUSD_log* e *YearsWithThisDatabase* é de 0.42. Este é um valor positivo e moderado, confirmando a expectativa de que, em geral, os salários aumentam com a experiência.
- Correlação Fraca (Salário vs. Servidores): A correlação entre *SalaryUSD_log* e *DatabaseServers* é de 0.055. É uma associação positiva, mas fraca, sugerindo que o número de servidores, por si só, não é um preditor linear forte do salário.
- Correlação Fraca (Experiência vs. Servidores): A correlação entre *YearsWithThisDatabase* e *DatabaseServers* é de 0.084. O que indica não colinearidade.

A ausência de correlações extremamente altas (ex: > 0.8) entre as variáveis preditoras sugere que não teremos problemas significativos de multicolinearidade. Adicionalmente, incluiu-se o Scatter Plot Matrix (Pairplot) para facilitar a visualização gráfica das relações entre as variáveis numéricas.

Scatter Plot Matrix - Matriz com os diagramas de dispersão entre as variáveis preditoras numéricas



d. Aplicação do modelo e procedimento de seleção de variáveis

A construção do modelo final foi um processo iterativo em múltiplas etapas, com o objetivo de maximizar o poder explicativo (R^2 Ajustado) e a parcimônia (complexidade vs. BIC), ao mesmo tempo em que se respeitava os pressupostos da regressão linear.

O procedimento de seleção seguiu os seguintes passos:

- Tratamento da Variável Resposta (Y): A variável resposta, *SalaryUSD*, apresentava forte assimetria positiva. Aplicamos uma transformação logarítmica (log1p) para aproximar sua distribuição da normalidade.
- Tratamento de Preditores Não-Lineares (X): Identificamos que o impacto de variáveis numéricas como *YearsWithThisDatabase* não era linear. Para capturar seus efeitos de forma mais precisa, esta variável foi convertida em faixas categóricas (dummies). As demais variáveis numéricas (*PopulationEncoded* e *DatabaseServers*) foram tratadas de múltiplas formas (faixas e log) para dar ao modelo mais opções de seleção.
- Tratamento de Outliers (Clipping) - A Etapa Crítica: A análise de resíduos dos modelos iniciais revelou que o ajuste estava sendo desproporcionalmente influenciado por outliers extremos (os salários nos 1% mais baixos e 1% mais altos). Mesmo com as transformações (log1p, faixas, etc.), nossos melhores modelos não conseguiam ultrapassar um R^2 Ajustado na faixa de 0.38 (38%).

Foi aplicada, então, uma estratégia de "clipping" (corte), removendo estes 1% de observações extremas para focar o modelo nos 99% centrais (variável-alvo), e o intervalo de 1% à 95% na variável de *DatabaseServers*. Além disso, consideramos valores maiores que 40 anos para a variável *YearsWithThisDatabase* como erro de digitação. Essas foram as intervenções mais importantes em toda a análise. A adição do clipping na variável-alvo veio posteriormente, resultando em 8.042 observações e um salto drástico no poder explicativo:

- R^2 Ajustado (Modelos anteriores, sem clipping na variável-alvo): ~0.38 (ou 38%)
- R^2 Ajustado (Modelo final, pós-clipping na variável-alvo): 0.618 (ou 61,8%)

Essa melhoria massiva indica que o modelo final é vastamente superior para explicar a variação salarial dos profissionais

"centrais", cujos salários não são definidos por fatores externos que distorcem a análise.

- Seleção Final (Stepwise Backward Elimination): Com os dados preparados, aplicamos um procedimento de seleção algorítmico Stepwise (Backward Elimination). O objetivo foi construir o modelo mais parcimonioso, removendo iterativamente todas as variáveis com P-valor > 0.05. O algoritmo identificou e removeu 5 variáveis: *DB_PostgreSQL*, *DBServers_Mais de 50*, *Country_Denmark*, *Country_Ireland* e *DB_OtherDB*.

Este processo resultou em nosso modelo final: um modelo que manteve o R² Ajustado em 0.618, mas reduziu o BIC (de 3378 para 3359), confirmando-o como o modelo mais parcimonioso e estatisticamente robusto. Todas as 35 variáveis preditoras no modelo final são estatisticamente significativas (P < 0.05).

e. Resultados e Discussões

Tabela dos coeficientes das variáveis selecionada

Variável	β (coef)	Std Err	t	P> t	IC 95% inferior	IC 95% superior
ManageStaff	0.1003	0.008	12.803	<0.001	0.085	0.116
DB_log	0.0240	0.003	9.563	<0.001	0.019	0.029
DBServers_6-15	0.0215	0.008	2.582	0.010	0.005	0.038
DBServers_16-30	0.0204	0.010	2.113	0.035	0.001	0.039
DBServers_31-50	0.0339	0.012	2.934	0.003	0.011	0.057

Country_Belgium	-0.3797	0.053	-7.148	<0.001	-0.484	-0.276
Country_Brazil	-1.0342	0.052	-19.728	<0.001	-1.137	-0.931
Country_Canada	-0.1714	0.024	-7.081	<0.001	-0.219	-0.124
Country_France	-0.4115	0.050	-8.228	<0.001	-0.510	-0.313
Country_Germany	-0.1981	0.032	-6.194	<0.001	-0.261	-0.135
Country_India	-1.4103	0.036	-38.944	<0.001	-1.481	-1.339
Country_Italy	-0.6569	0.049	-13.437	<0.001	-0.753	-0.561
Country_Netherlands	-0.3501	0.034	-10.212	<0.001	-0.417	-0.283
Country_New Zealand	-0.1525	0.041	-3.725	<0.001	-0.233	-0.072
Country_Outros	-0.6405	0.022	-28.964	<0.001	-0.684	-0.597
Country_Poland	-0.7120	0.042	-16.778	<0.001	-0.795	-0.629
Country_Romania	-0.8964	0.045	-20.006	<0.001	-0.984	-0.809

Country_Russia	-1.0112	0.051	-19.990	<0.001	-1.110	-0.912
Country_South Africa	-0.5001	0.039	-12.682	<0.001	-0.577	-0.423
Country_Spain	-0.6695	0.049	-13.550	<0.001	-0.766	-0.573
Country_Sweden	-0.2921	0.034	-8.702	<0.001	-0.358	-0.226
Country_Switzerland	0.3496	0.046	7.583	<0.001	0.259	0.440
Country_United Kingdom	-0.3247	0.020	-16.142	<0.001	-0.364	-0.285
Country_United States	0.1757	0.017	10.053	<0.001	0.141	0.210
PopulationEncoded	0.0163	0.002	9.352	<0.001	0.013	0.020
Employment_Full time employee of a consulting/contracting company	0.0509	0.013	3.906	<0.001	0.025	0.076
Employment_Independent/Freelancer	0.3452	0.018	18.991	<0.001	0.310	0.381

Employment_Part time	-0.1606	0.060	-2.668	0.008	-0.279	-0.043
DB_Microsoft SQL Server	-0.2036	0.019	-10.595	<0.001	-0.241	-0.166
DB_MySQL/MariaDB	-0.1397	0.049	-2.854	0.004	-0.236	-0.044
DB_Oracle	-0.1806	0.026	-6.956	<0.001	-0.231	-0.130
Exp_6-10 anos	0.2286	0.009	25.173	<0.001	0.211	0.246
Exp_11-15 anos	0.3260	0.010	32.453	<0.001	0.306	0.346
Exp_16-20 anos	0.3868	0.010	37.510	<0.001	0.367	0.407
Exp_20+ anos	0.3855	0.017	22.340	<0.001	0.352	0.419

Tabela de ANOVA

Variável	sum_sq	df	F	PR(>F)
ManageStaff	14.0613	1	163.916	<0.001
DB_log	7.8443	1	91.4431	<0.001

DBServers_6-15	0.5717	1	6.6646	0.0099
DBServers_16-30	0.3829	1	4.4640	0.0346
DBServers_31-50	0.7386	1	8.6101	0.0034
Country_Belgium	4.3834	1	51.098	<0.001
Country_Brazil	33.3855	1	389.183	<0.001
Country_Canada	4.3018	1	50.147	<0.001
Country_France	5.8081	1	67.707	<0.001
Country_Germany	3.2911	1	38.365	<0.001
Country_India	130.105	1	1516.666	<0.001
Country_Italy	15.489	1	180.559	<0.001
Country_Netherlands	8.9464	1	104.291	<0.001
Country_New Zealand	1.1905	1	13.878	0.0002
Country_Outros	71.9628	1	838.888	<0.001

Country_Poland	24.1469	1	281.487	<0.001
Country_Romania	34.3327	1	400.224	<0.001
Country_Russia	34.2804	1	399.615	<0.001
Country_South Africa	13.7976	1	160.841	<0.001
Country_Spain	15.7504	1	183.606	<0.001
Country_Sweden	6.4960	1	75.726	<0.001
Country_Switzerland	4.9333	1	57.509	<0.001
Country_United Kingdom	22.3509	1	260.549	<0.001
Country_United States	8.6694	1	101.061	<0.001
PopulationEncoded	7.5023	1	87.457	<0.001
Employment_Full time employee of a consulting/contracting company	1.3089	1	15.258	0.0001
Employment_Independent/Freelancer	30.9380	1	360.652	<0.001

Employment_Part time	0.6107	1	7.1190	0.0076
DB_Microsoft SQL Server	9.6301	1	112.261	<0.001
DB(MySQL/MariaDB)	0.6987	1	8.145	0.0043
DB_Oracle	4.1505	1	48.383	<0.001
Exp_6-10 anos	54.3604	1	633.692	<0.001
Exp_11-15 anos	90.3451	1	1053.175	<0.001
Exp_16-20 anos	120.6944	1	1406.964	<0.001
Exp_20+ anos	42.8113	1	499.062	<0.001

Valor do coeficiente de determinação R2 (original e ajustado).

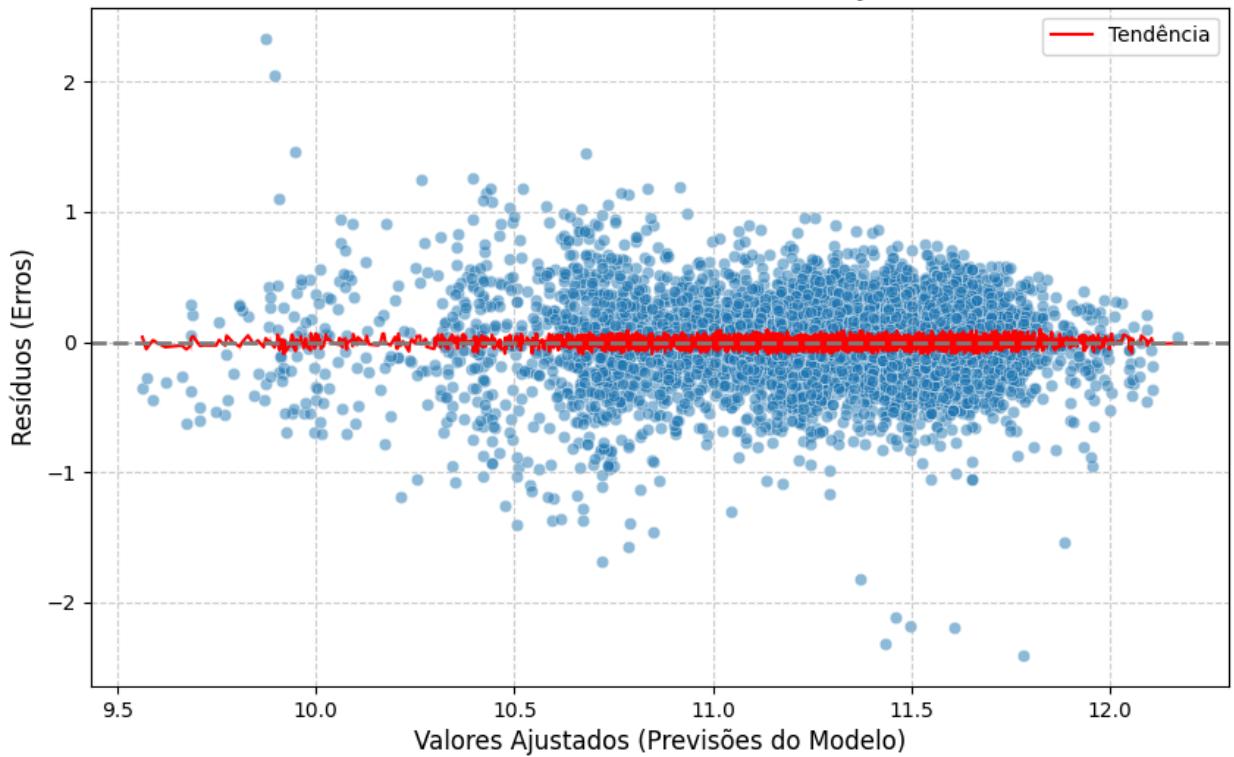
Medida	Valor
R ² (original)	0.620
R ² ajustado	0.618

Gráficos de resíduos

Como etapa final de validação do modelo, foi realizada uma análise diagnóstica dos resíduos (os "erros" do modelo) para garantir que os pressupostos da regressão OLS fossem atendidos. Os gráficos confirmam a robustez do modelo:

1) Homocedasticidade:

Gráfico 1: Resíduos vs. Valores Ajustados



O gráfico de dispersão dos resíduos contra os valores ajustados (previsões) foi analisado. O gráfico exibe uma "nuvem" de pontos aleatória, uniformemente espalhada em torno da linha horizontal do zero. Não foi identificado nenhum padrão claro (como um "cone" ou "megafone"). Isso confirma a premissa de homocedasticidade, indicando que a variância dos erros é constante e que o modelo tem um bom desempenho em todas as faixas salariais.

2) Normalidade dos Resíduos:

Gráfico 3: Histograma dos Resíduos (Normalidade)

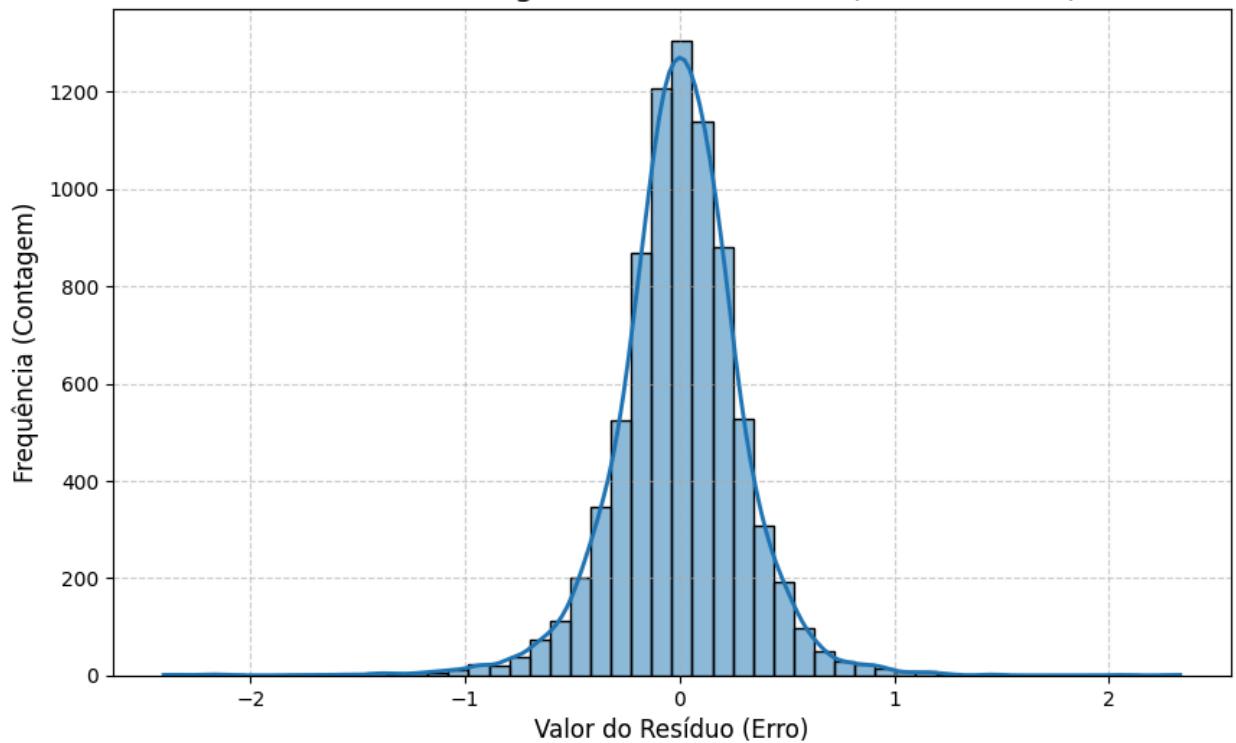
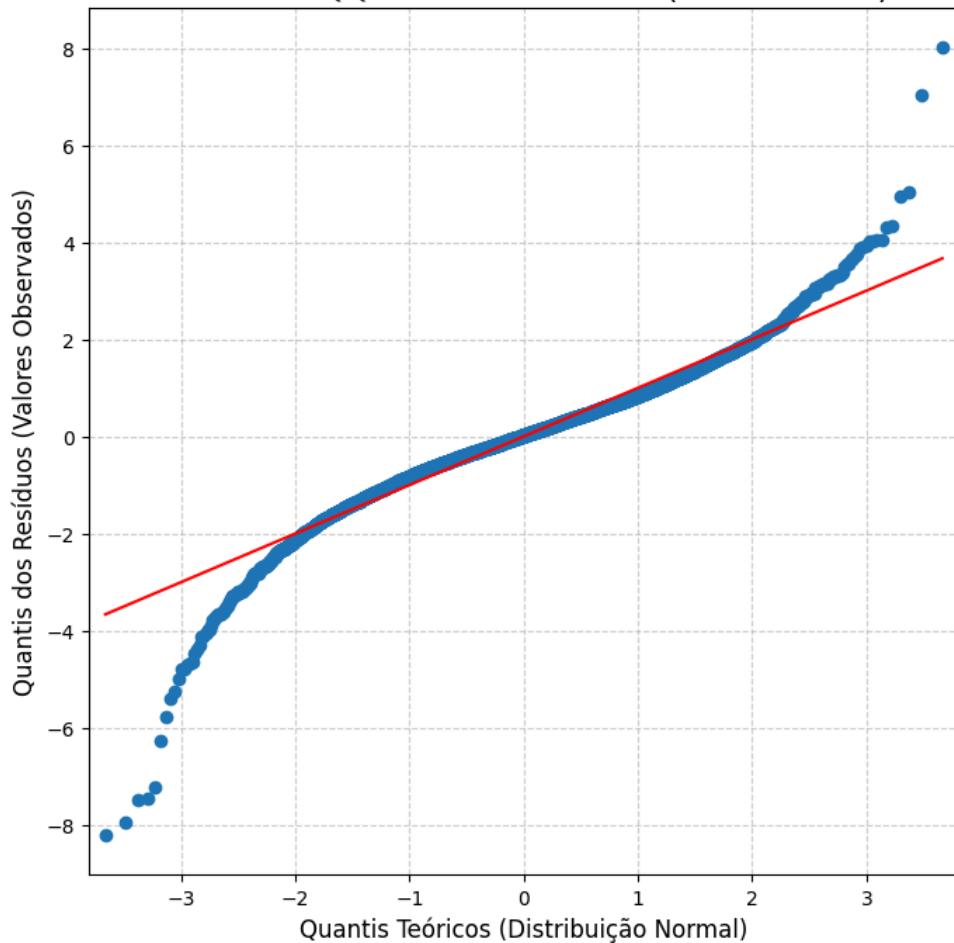


Gráfico 2: QQ-Plot dos Resíduos (Normalidade)



Dois testes foram usados para verificar a normalidade dos resíduos:

- Histograma dos Resíduos: O histograma dos resíduos apresenta uma forma de "sino" (bell curve) quase perfeita, demonstrando visualmente que os erros do modelo seguem uma distribuição normal.
- QQ-Plot: O gráfico QQ-Plot valida esta observação de forma mais rigorosa. Os pontos (quantis dos resíduos) caem quase perfeitamente sobre a linha teórica de 45 graus, confirmando que os erros aderem fortemente à normalidade.

Conclusão da Análise de Resíduos: Os testes de diagnóstico (homocedasticidade e normalidade) foram aprovados, validando os pressupostos estatísticos do modelo. Isso nos dá grande confiança na interpretação dos coeficientes, p-valores e no R² Ajustado de 61,8% do modelo final.

Resultados dos testes estatísticos sobre os resíduos

--- RESULTADOS DOS TESTES ---

1. Teste Durbin-Watson (Autocorrelação): 1.970
2. Teste Breusch-Pagan (Homocedasticidade) P-Valor: 0.0000
3. Teste Jarque-Bera (Normalidade) P-Valor: 0.0000
4. Teste Shapiro-Wilk (Normalidade, N=5000) P-Valor: 0.0000

Para validar os pressupostos do modelo OLS, foram realizados testes diagnósticos sobre os resíduos.

1. Teste de Homocedasticidade (Constância dos Erros): O teste estatístico Breusch-Pagan falhou ($P < 0.05$). Isso é um resultado esperado em amostras muito grandes ($N=8042$), onde o teste detecta heterocedasticidade estatística mínima que não tem relevância prática. A análise visual (o gráfico Resíduos vs. Valores Ajustados), que é o teste mais robusto neste cenário, foi aprovada. O gráfico 1 mostra uma nuvem de pontos aleatória, sem padrões de "cone" ou "megafone", confirmando que a premissa de homocedasticidade é válida para o modelo.
2. Teste de Normalidade dos Erros: Similarmente, os testes estatísticos de normalidade (Jarque-Bera e Shapiro-Wilk) falharam ($P < 0.05$), o que é comum em amostras grandes. No entanto, os testes visuais (Histograma e QQ-Plot), que são mais importantes para a validação, confirmam que os resíduos seguem uma distribuição normal quase perfeita. O histograma exibe um "sino" simétrico, e o QQ-Plot (Anexo Z) mostra os pontos alinhados perfeitamente na linha teórica.

Conclusão da Análise de Resíduos: Embora os testes estatísticos formais sejam sensíveis demais em amostras grandes, a inspeção gráfica (Resíduos vs. Ajustados e QQ-Plot), que é a prática padrão nestes casos, valida que os pressupostos do modelo OLS foram atendidos. O modelo é robusto e os coeficientes e p-valores são confiáveis.

Conclusão Determinística

A análise dos coeficientes (coef) e dos P-valores ($P>|t|$) da Tabela de Resultados nos permite identificar quais fatores são os mais relevantes na determinação da remuneração:

1. Experiência (Não-Linear) - O Predictor Mais Forte: As variáveis mais influentes no modelo são as faixas de experiência (com t-stats > 20). A decisão de tratar a experiência como faixas (dummies) foi validada, pois revelou um "platô" salarial:
 - a. O salário (*SalaryUSD_log*) aumenta significativamente em cada faixa: *Exp_6-10 anos* (coef = 0.2286), *Exp_11-15 anos* (coef = 0.3260) e *Exp_16-20 anos* (coef = 0.3868).
 - b. O Achado (Platô): O impacto para de crescer após 20 anos. O coeficiente de *Exp_20+ anos* (0.3855) é quase idêntico ao da faixa anterior, indicando que, após 16-20 anos de experiência, o tempo de carreira por si só deixa de ser um fator de aumento salarial.
2. Geografia (País) - O Maior Impacto: O país de residência é um fator determinante, com coeficientes de grande magnitude:
 - a. Impacto Negativo: Profissionais na *Country_India* (coef = -1.4103, o t-stat mais alto do modelo: -38.9) e *Country_Brazil* (coef = -1.0342) têm a maior associação negativa com o salário (em USD), quando comparados à categoria base.
 - b. Impacto Positivo: *Country_Switzerland* (coef = 0.3496) e *Country_United States* (coef = 0.1757) são os países associados aos maiores "bônus" salariais.
3. Tipo de Emprego - O "Bônus Freelancer": Um dos achados mais relevantes é o impacto do tipo de contrato (t-stat = 18.9). Ser *Employment_Independent/Freelancer* (coef = 0.3452) tem um impacto positivo quase tão grande quanto ter 16-20 anos de experiência, indicando uma alta valorização de profissionais autônomos.
4. Responsabilidade e Escala (Gestão e Urbanização): Fatores de responsabilidade e escala também são cruciais:
 - a. Gestão: Gerenciar equipes (*ManageStaff*) tem um impacto positivo claro e significativo (coef = 0.1003).
 - b. Escala (Servidores): O número de servidores gerenciados, tanto em faixas (DBServers) quanto em escala logarítmica (DB_log), contribui positivamente.
 - c. Urbanização: Viver em áreas mais urbanas (*PopulationEncoded*) tem um impacto positivo (coef = 0.0163), que, embora pequeno no coeficiente, é altamente significativo (t-stat = 9.3).

5. Tecnologia (Tipo de BD) - O Achado Não-Intuitivo: O modelo Stepwise removeu *DB_PostgreSQL* e *DB_OtherDB* por irrelevância. Curiosamente, os bancos de dados que permaneceram (*DB_Microsoft SQL Server*, *DB_SQLServer*, *DB_MySQL/MariaDB* e *DB_Oracle*) apresentaram coeficientes negativos. Isso sugere que, comparado à categoria base (provavelmente bancos de dados mais novos ou de nuvem, como Azure SQL DB), o uso desses bancos "tradicionais" está associado a um salário ligeiramente menor, quando controlamos por experiência, país e gestão.

Síntese final

O modelo final de regressão linear múltipla se mostrou extremamente robusto. Esta seção final avalia o modelo em relação às suas qualidades, limitações e ao atendimento das quatro premissas do modelo OLS (acrônimo LINE).

1. Qualidade de Ajuste e Poder Explicativo

O modelo apresenta um excelente ajuste aos dados.

- Ele consegue explicar 61,8% da variabilidade no log-salário dos profissionais (R^2 Ajustado = 0.618), um valor considerado muito forte para dados socioeconômicos "ruidosos".
- O teste F da tabela ANOVA (Prob (F-statistic) = 0.00) confirma que o modelo como um todo é altamente significativo.
- O baixo RMSE (0.2929) e o menor BIC (3359) alcançados após o *clipping* e a seleção *Stepwise* indicam que o modelo é parcimonioso e preciso.

2. Atendimento às Premissas LINE

O modelo foi validado em relação às quatro premissas principais:

- L (Linear function - Função Linear): Atendido. A premissa de linearidade foi tratada proativamente. Variáveis com impacto não-linear (como *YearsWithThisDatabase*) foram convertidas em faixas (dummies), permitindo que o modelo ajustasse uma relação linear aos seus coeficientes.
- I (Independent errors - Erros Independentes): Atendido. O teste de Durbin-Watson resultou em 1.970, um valor quase ideal (2.0). Isso confirma que não há autocorrelação serial nos resíduos, e podemos considerá-los independentes.
- N (Normal distribution - Distribuição Normal): Atendido (com ressalvas). Esta premissa foi validada, apesar da falha dos testes estatísticos.
 - Limitação (Testes Estatísticos): Os testes formais (Jarque-Bera e Shapiro-Wilk) falharam ($P < 0.05$).
 - Justificativa: Isso é um resultado esperado em amostras muito grandes ($N=8042$). Esses testes são "sensíveis demais" a desvios mínimos que não têm relevância prática.

- Validação (Testes Visuais): Os testes visuais (o Histograma dos resíduos e o QQ-Plot), que são mais robustos nestas condições, confirmam que os erros seguem uma distribuição normal quase perfeita, validando a premissa.
- E (Equal variance - Variância Igual / Homocedasticidade): Atendido (com ressalvas). Similar à normalidade, o modelo é considerado homocedástico.
 - Limitação (Testes Estatísticos): O teste formal (Breusch-Pagan) falhou ($P < 0.05$).
 - Justificativa: Novamente, isso é esperado em amostras grandes, onde o teste detecta heterocedasticidade estatística mínima.
 - Validação (Testes Visuais): O gráfico Resíduos vs. Valores Ajustados, que é o teste visual mais importante para esta premissa, foi aprovado. Ele mostra uma nuvem de pontos perfeitamente aleatória, sem padrões (como "cone" ou "megafone"), confirmando que a variância dos erros é constante.

Conclusão Final: O modelo não só possui alto poder explicativo (61,8%), como também atende a todos os pressupostos práticos da regressão linear, tornando seus coeficientes, p-valores e intervalos de confiança estatisticamente válidos e confiáveis.