جامعة دمشق
كلية الهندسة المعلوماتية
قسم الذكاء الصنعي

استكشاف المعرفة & التعلم التلقائي

توقع رواتب محللي البيانات

Predicting Data Analyst Salaries

نابغ صايغ – فرح صايغ – مناف صعوب – ماريا رزق – ميشيل غيث – بشر خميسي – مجد القائد

الخلاصة:

في الوقت الحاضر، زاد الطلب على محللي البيانات بشكل كبير. أولاً، قمنا بتنزيل بيانات محللي البيانات من Kaggle، ثم قمنا بتحليل معلومات التوظيف عن طريق طريقة استخراج النص مثل إحصائيات تكرار الكلمات وتكنولوجيا التصور, ثم قمنا بتلخيص المهارات المطلوبة لمحلل البيانات. لقد وجدنا أنه بالمقارنة مع SQL، تميل الشركة إلى استخدام Excel لتحليل البيانات، وتختلف متطلبات التوظيف لكل نوع من أنواع الشركات. ثم استخدمنا تقنيات Machine Learning للتنبؤ بمتوسط الراتب للوظيفة المعلن عنها.

1) المقدمة:

محلل البيانات هو محترف يقوم بجمع وتحليل البيانات عبر الشركة لاتخاذ قرارات مستنيرة ومساعدة أعضاء الفريق الاخرين والقيادة في اتخاذ القرارات السليمة ويشمل دوره على جميع البيانات واستردادها من مصادر محددة بالاضافة الى عملية التحليل البيانات باستخدام ادوات لاستخراج المعلومات المفيدة وتكوين الرؤى ثم تحويل البيانات الى معلومات واضحة وتقييمها من حيث الجودة والفائدة .

يحتاج الى ان يملك مهارات اساسية مثل القدرة على جمع البيانات وتحديد المصادر , استخدام الادوات المناسبة لتحليل البيانات مثل: PYTHON, SQL وغيرها ومن المهارات المهمة هي طريقة تصور البيانات وتوضيحها في تصورات بيانية وتقارير توضيحية باستخدام ادوات مثل: Power BI, Tableau

هدفنا النهائي من هذا البحث هو فهم المهارات المطلوبة من محللي البيانات والتنبؤ بمتوسط رواتبهم .

2) وصف البيانات:

يتم استخراج البيانات الموجودة في مجموعة البيانات من موقع Kaggle ، والبيانات مستخلصة من موقع Glassdoor وهو موقع ويب لنشر الوظائف. تحتوي مجموعة البيانات على بيانات تتعلق بوظائف تحليل البيانات والرواتب وغير ذلك الكثير، مما يوفر رؤية واضحة لفرص العمل. إنها مليئة بالتفاصيل الأساسية مثل المسميات الوظيفية والرواتب المقدرة والأوصاف الوظيفية وتقييمات الشركة ومعلومات الشركة الرئيسية مثل الموقع والحجم والصناعة.

Field	Type
Job title	object
Salary estimate	object
Job description	object
Rating	numeric
Company name	object
Location	object
Headquarters	object
Size	object
Founded	numeric
Type of ownership	object
Industry	object
Sector	object
Revenue	object
Competitors	object
Easy apply	object

نلاحظ من الجدول السابق التفصيل المتوفر في الداتا سيت حيث يتم تحديد المسمى الوظيفي والراتب المتوقع ووصف الوظيفة وهي قيم نصية حيث يتم تحديد الراتب على شكل مجال. ونرى ايضا اسم الشركة وموقعها اضافة الى موقع الوظيفة وحجمها من حيث عدد الموظفين وهي قيمة مقسمة وموزعة لذلك تصنف كغرض. بالاضافة الى نوعية الشركة من الناحية الادارية ومجالات عملها والقسم الذي سيتم العمل فيها ونرى ايذا العائدات والمنافسين كل المذكورة قيم نصية مفصلة ومحددة, اما القيم الرقمية فهي تقييم الشركة العام وتاريخ انشائها.

3. الاعمال ذات الصلة:

- اقترحت العديد من الدراسات السابقة أساليب لدراسة وتقييم الفجوة بين العرض والطلب في المهن المتعلقة بتحليل البيانات. إحدى المهارات المطلوبة لاستخراج متطلبات العمل من ال job description باستخدام word2vec verctorizer لاستخلاص أكثر الكلمات المكررة لل job description , ثم دراسة التشابه بين هذه الكلمات. [1]
- هنا كان المؤلف يركز أكثر على كاليفورنيا. قام بتحليل البيانات الاستكشافية (EDA) على مجموعة البيانات وسبقها إجراء تحليل الانحدار ووجد أن :لا تزال الاختلافات الإقليمية (موقع العمل) هي العامل الأكثر حسماً في تباين الرواتب. لا تدفع شركات تكنولوجيا المعلومات بالضرورة أجوراً أعلى، ولكن شركات تكنولوجيا المعلومات في كاليفورنيا تفعل ذلك .(CA_IT) تميل الشركات الصغيرة (ذات الإيرادات من 1 إلى 5 ملايين دولار أمريكي) إلى دفع أجور أعلى.تميل الشركات التي يقع مقرها في نيوجيرسي إلى دفع أجور أعلى(NJ_HQ) . يحصل المحللون الذين لديهم خبرة في MYSQL على أجور أعلى. [2]
- أجرت دراسة اخرى تحليلًا استكشافيًا للبيانات ووجدت أن أهم المهارات المطلوبة لمحللي البيانات هي SQL وExcel .كما وجدت أن الوظائف التي لا تذكر درجة علمية (بكالوريوس أو ماجستير) لها متوسط رواتب يبلغ 73 ألف دولار، بينما الوظائف التي تذكر درجة علمية (بكالوريوس أو ماجستير) لها متوسط رواتب يبلغ 71 ألف دولار، مما يشير إلى أن الدرجات العلمية لم تكن لها فارق كبير في الرواتب. [3]
- دراسة اخرى كان هدفها هو استخدام تقنيات التعلم الآلي لتحليل اتجاهات الرواتب داخل صناعة علوم البيانات. في البداية، قدمت هذه الدراسة نظرة عامة على أربعة نماذج من التعلم الآلي: Random Forests، extreme Gradient Boosting (XGBoost)، والشبكات العصبية، وانحدار المتجه الداعم(SVR) ، موضحةً مبادئها الأساسية وخصائصها. بعد ذلك، جمعت هذه الدراسة البيانات وقامت بمعالجتها مسبقًا وشاركت في هندسة الميزات مع بيانات الرواتب من قطاع علوم البيانات. ثم تم استخدام هذه النماذج الأربعة للتعلم الآلي للتنبؤ بالرواتب، وتم فحص نتائج النماذج الناتجة بعناية فائقة. من خلال إجراء تحليل مقارن وتقييم أداء كل نموذج، تم تحديد نقاط القوة والضعف لكل منها. الاستنتاج الرئيسي هو أن XGBoost يؤدي أفضل في التنبؤ بالرواتب، بينما تكون الشبكات العصبية أكثر دقة وتعقيدًا، ولكن لانحدار المتجه الداعم (SVR) تطبيقات محدودة . [4]

- هذه الورقة البحثية تهدف إلى مراجعة المنهجيات الحديثة والقائمة لبناء نموذج تنبؤ بالرواتب أكثر ملاءمة يعتمد على المهارات المتخصصة والمزايا الوظيفية المعطاة في مجال علوم البيانات. كما تشمل عملية اكتشاف المعرفة تحديد المشكلات القائمة في موارد البشرية في مجال علوم البيانات وأكثر المهارات المطلوبة للاستكشاف المبكر وتحديد المتغيرات الداخلية. نظرًا لأن علوم البيانات تشمل بُعدًا كبيرًا من المناصب والمسؤوليات، تم تصميم مجموعة البيانات التجريبية لتشمل عوامل تعتمد على المهارات والمزايا الوظيفية لتنبؤات الرواتب أكثر دقة. تم تصنيف منهجيات التعلم الآلي المعيارية المراجعة للمشكلات ذات الصلة إلى ثلاث فئات رئيسية مع نقاط قوة فردية تحت ظروف ومتطلبات مختلفة. تعتبر الطرق الإحصائية أفضل في تقديم علاقات المتغيرات مع إمكانية ضبط المعاملات الاستثنائية إذا كانت الخطية موجودة. تعتبر طرق التعلم الآلي المجمعة مثل الغابات العشوائية التي تجمع بين مصنفات متعددة للتنبؤ أكثر استقرارًا ودقة. تتمتع الشبكات العصبية المبنية على التعلم العميق بتخصص قوي في التعامل مع البيانات غير المصنفة وتعديلات الإطار. علاوة على ذلك، تم إدراك أن مجموعات البيانات الضخمة مع المتغيرات المناسبة وطريقة ضبط البحث الشبكي تحقق أداءً أكبر وأكثر موثوقية . [5]
- وفي هذا البحث الاخير تم استخدام تحليل المكونات الرئيسية (PCA) لمساعدة معالجة البيانات الديموغرافية واستخراج المعلومات ذات الصلة التي لها تأثير كبير على نموذج التنبؤ. وعليه، للتنبؤ بالرواتب، استخدمت البحث نهج شبكة عصبية عميقة قوية (PCA-DNN) لزيادة أداء التنبؤ على مصنفات التعلم الآلي القائمة مثل الأشجار القرارية (DT) والغابات العشوائية .(RF). تم الحصول على دقة تصنيف أفضل باستخدام خوارزميات-PCA DNNالمقترحة من حيث الدقة والتحديد والاسترجاع ومعدل .F بالإضافة إلى ذلك، حقق نموذج DNN المقترح أداء تنبؤ أعلى بنسبة 94.9٪ مقارنةً بـ DT و RF التي حققت درجة MAE بنسبة 89.6٪ و 76.4٪ على التوالي. يشير نتيجة النتائج إلى أن النموذج المقترح لديه خطأ في التنبؤ بنسبة 5.1٪، وهو أقل مقارنةً بـ DT و RFمع خطأ في التنبؤ بنسبة 10.4 و 23.6٪ على التوالي. هذا يؤسس لتفوق خوارزمية التعلم العميق على خوارزميات التعلم الآلي التقليدية في مهمة تصنيف وتنبؤ الرواتب. [6]

4. المنهجية المتبعة:

يصف هذا القسم العملية التي تم اتباعها لبناء مجموعة البيانات ويقدم أيضًا تحليلًا استكشافيًا موجزًا للبيانات يسلط الضوء على بعض القضايا ذات الصلة التي قد تساعد الباحثين الآخرين في وضع أيديهم بسرعة على مجموعة البيانات والعمل معها، مثل الطبيعة غير المتوازنة للبيانات.

i) أجرينا تحليلًا مختصرًا للبيانات الاستكشافية في Python 3 باستخدام المكاتب التالية:

- Numpy	- Seaborn	- Mlxtend
- Pandas	- Wordcloud	- sklearn
- Matplotlib	- apriori	

ii) التحليل الوصفي وملء القيم المفقودة:

نقوم أولاً بتحميل بياناتنا وعرض معلوماتها. نرى أن لدينا 2253 صفًا و16 ميزة

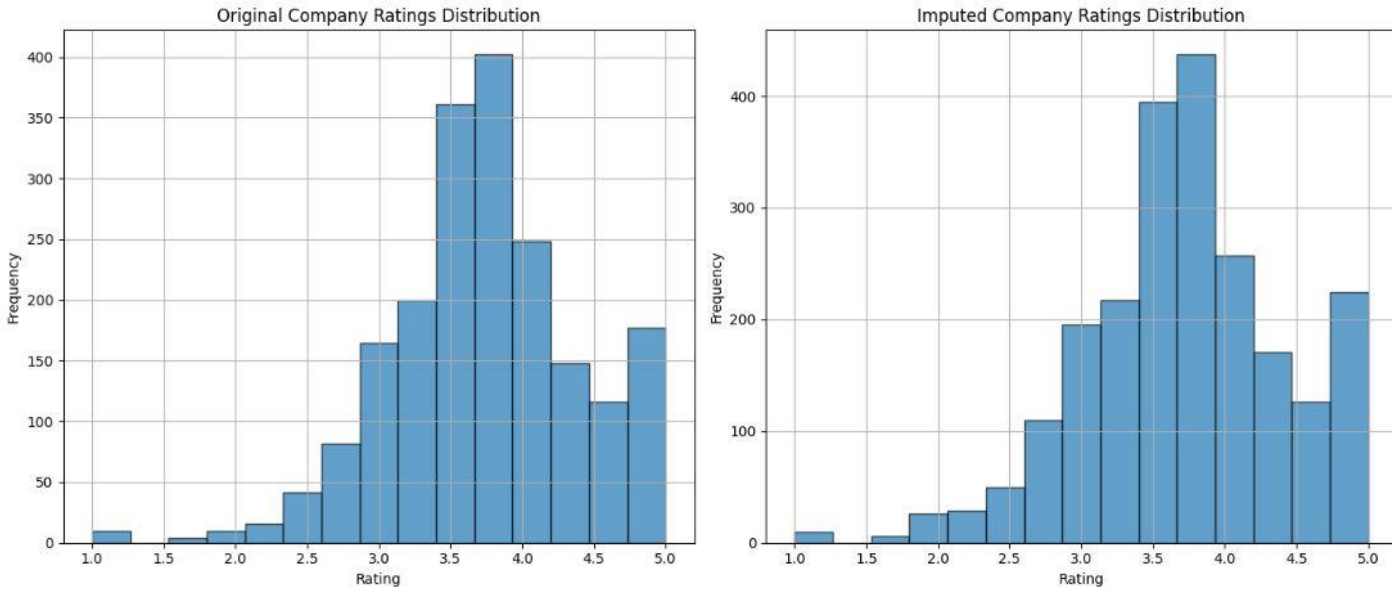
في النظرة الأولى، نرى أنه لا توجد قيم مفقودة باستثناء واحدة في اسم الشركة، ولكن بعد مزيد من الاستكشاف نرى أن هناك قيمًا في البيانات إما -1 أو -1.0 أو 1 إما بتنسيق سلسلة أو أرقام. يمكن أن تشير هذه القيم إلى القيم المفقودة في بياناتنا. لذلك نستبدل هذه القيم بـ NaN ثم نرى القيم المفقودة الفعلية.

	missing_val	missing_val_ratio	dtypes
Unnamed: 0	0	0.000000	int64
Job Title	0	0.000000	object
Salary Estimate	0	0.000000	object
Job Description	0	0.000000	object
Rating	0	0.000000	float64
Company Name	1	0.044385	object
Location	0	0.000000	object
Headquarters	0	0.000000	object
Size	0	0.000000	object
Founded	0	0.000000	int64
Type of ownership	0	0.000000	object
Industry	0	0.000000	object
Sector	0	0.000000	object
Revenue	0	0.000000	object
Competitors	0	0.000000	object
Easy Apply	0	0.000000	object

	missing_val	missing_val_ratio	dtypes
Unnamed: 0	0	0.000000	int64
Job Title	0	0.000000	object
Salary Estimate	1	0.044385	object
Job Description	0	0.000000	object
Rating	272	12.072792	float64
Company Name	2	0.088771	object
Location	0	0.000000	object
Headquarters	172	7.634265	object
Size	163	7.234798	object
Founded	660	29.294274	float64
Type of ownership	163	7.234798	object
Industry	353	15.667998	object
Sector	353	15.667998	object
Revenue	163	7.234798	object
Competitors	1732	76.875277	object
Easy Apply	2173	96.449179	object

ثم نقوم بإسقاط الأعمدة غير الضرورية: [غير مسمى: 0، المنافسون، سهل التقديم، الصناعة] لأننا لن نستخدمها. بعد ذلك نقوم بتغيير أسماء الأعمدة لتسهيل الاستخدام .نلقي نظرة على job_title ونرى أنه لا يحتوي على قيم مفقودة ولكن لديه 1272 قيمة فريدة، لذلك نبدأ بتوحيد القيم يدويًا، وننتهي بـ 833 قيمة فريدة. ثم ننظر إلى الرواتب التقديرية فنرى أن بها قيمة واحدة مفقودة فقط، فنقوم بملئها من خلال النظر إلى نفس اسم الشركة وملئها بنفس القيمة التي تقدمها الشركة. ثم نقوم بتغيير القيم من المظهر بهذا الشكل ((37K-66K(Glassdoor est) الى هذا الشكل:

(66- 37) لتسهيل الاستخدام. أما بالنسبة للتقييم فنرى أن لدينا 272 قيمة مفقودة وباقي القيم تتراوح بين 1-5، وهنا نقوم بملء القيم المفقودة لكل شركة فريدة باستخدام توزيع جاما للحفاظ على التوزيع كما هو.



بالانتقال إلى اسم الشركة، نرى أن لدينا قيمتين مفقودتين هنا، حيث نقوم فقط بإزالة هذين الصفين من بياناتنا، كما نقوم أيضًا بتنظيف اسم الشركة من التصنيف الموجود فيه.

أما بالنسبة للموقع، فنصحح القيمة (Greenwood Village, Arapahoe, CO) لتصبح (Greenwood Village, CO). بالنسبة للمقر الرئيسي لدينا قيم تحتاج إلى تصحيح كما هو الحال في الموقع، فنرى أن لدينا 172 قيمة مفقودة، وهنا نفترض أن المقر الرئيسي للشركة هو نفس الموقع إذا كان المقر الرئيسي مفقودًا ونملأه بناءً على ذلك. ننتقل إلى المؤسسة هنا لدينا 660 قيمة مفقودة، نحاول أولاً البحث عن شركة لها نفس الاسم وتم ملء عمود المؤسسة وإذا لم ينجح ذلك، فإننا نملأ القيم المفقودة بـ -1. أما بالنسبة للحجم الذي يحتوي على 163 قيمة مفقودة، ونوع_الملكية 163 قيمة، والقطاع الذي يحتوي على 353 قيمة مفقودة، فإننا نملأ هذه القيم المفقودة بـ (غير معروف). أخيرًا، تحتوي الإيرادات على 163 قيمة مفقودة وسنملأها بـ (غير معروف / غير قابل للتطبيق).

iii) هندسة السمات:

ننتقل الآن إلى هندسة الميزات لاستخراج ميزات جديدة من بياناتنا

أولاً، نستخرج الحد الأدنى للرواتب، الحد الأقصى للرواتب، متوسط الراتب من تقدير الراتب، المتوسط هو نتيجة (الحد الأدنى + الحد الأقصى) مقسومًا على 2. ثم نقوم بعمل خانات فئوية لمتوسط الراتب [0، 40، 70، np.inf] بالقيم [منخفض، متوسط، مرتفع].

ثم نستخرج المهارات التي نريدها: ["SQL"، 'Excel'، 'Tableau'، R، 'Power BI'، 'SAS'، 'Word'، 'PowerPoint'، 'AWS'، 'Python'، 'Master', 'Bachelor"] من job_description, ثم نقوم بإنشاء عمود يسمى No_degree ونملأه بالرقم 1 إذا كان كل من البكالوريوس والماجستير 0. بعد ذلك نستخرج من الموقع عمودين job_city, job_state. ونفعل الشيء نفسه بالنسبة للمقر الرئيسي ونسمي الأعمدة hq_city، hq_state

وأخيرا نقوم بعمل عمود العمر من العمود المؤسس وذلك بطرح 2024 من القيمة الموجودة وإذا كانت القيم -1 نحتفظ بها -1. ثم نقوم بإنشاء عمود تصنيفي يسمى age_bin للعمر [-1, 0, 25, 75, np.inf] مع القيم ['Unknown', 'new', 'old', 'very old'].

iv) Outlier detection:

أولاً نستخدم boxplot على الأعمدة الرقمية لرؤية البيانات

ثم نستخدم طريقة IQR لتحديد البيانات الخارجية وإحصائها:

rating	42
founded	0
min_salary	86
max_salary	82
average_salary	139
SQL	0
Excel	0
Tableau	0
R	441
Power BI	180

SAS	411
Word	272
PowerPoint	194
AWS	275
Python	0
Bachelor	0
Master	498
No_degree	0
age	186
dtype: int64	

لقد اكتشفنا عددًا من القيم المتطرفة في بياناتنا ولكننا لن نفعل أي شيء لها لأن إزالتها قد تؤثر على التباين الطبيعي في البيانات

v) التحليلات:

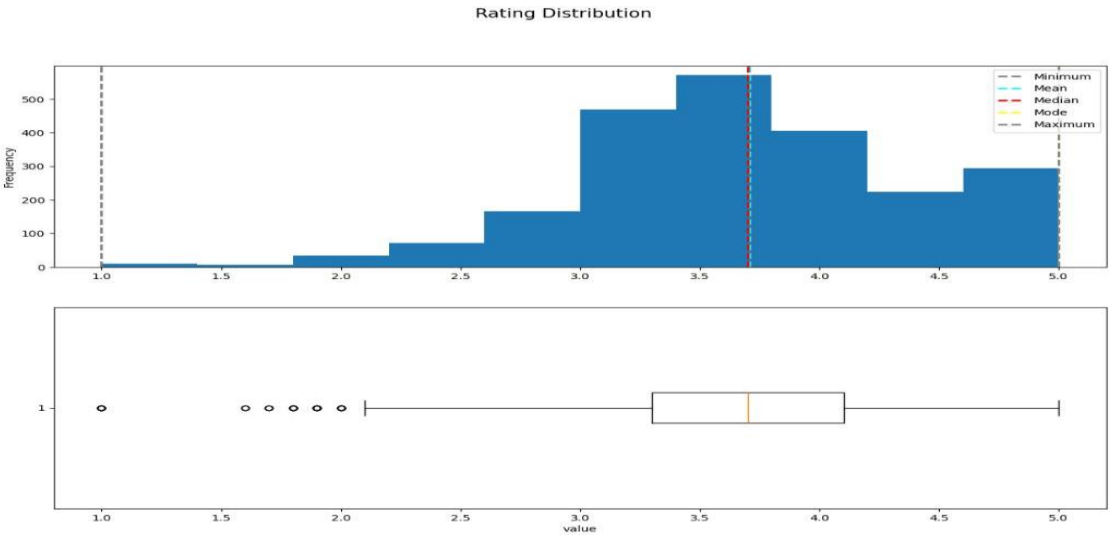
(1) نبدأ بالتحليل أحادي المتغير على قيمنا العددية ثم على القيم الفئوية:

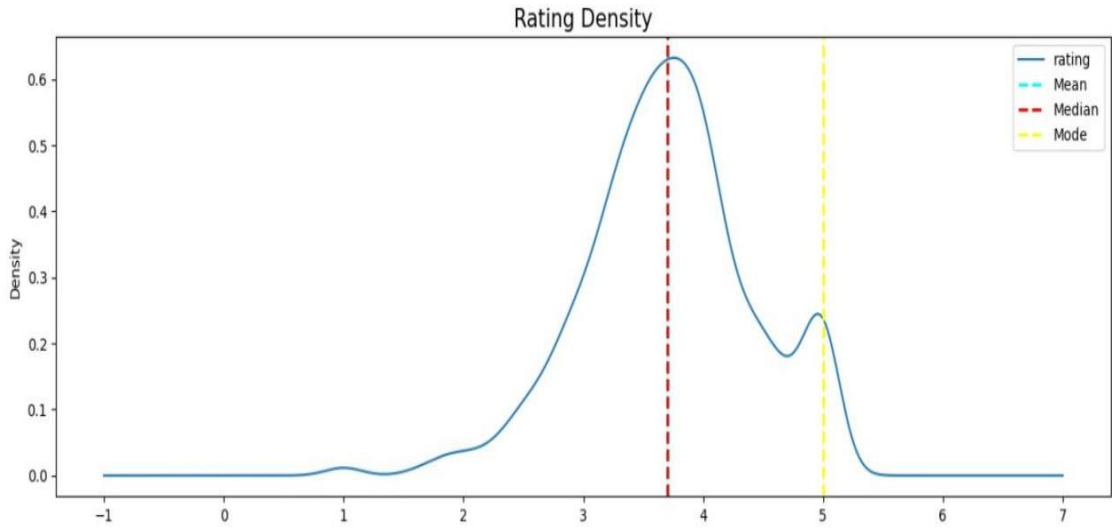
A. القيم العددية:

	rating	founded	min_salary	max_salary	average_salary	SQL
count	2251.000000	2251.000000	2251.000000	2251.000000	2251.000000	2251.000000
mean	3.709596	1404.204354	54.278543	89.996446	72.137494	0.616171
std	0.710115	899.371987	19.573024	29.312053	23.597327	0.486425
min	1.000000	-1.000000	24.000000	38.000000	33.500000	0.000000
25%	3.300000	-1.000000	41.000000	70.000000	58.000000	0.000000
50%	3.700000	1980.000000	50.000000	87.000000	69.000000	1.000000
75%	4.100000	2002.000000	64.000000	104.000000	80.500000	1.000000
max	5.000000	2019.000000	113.000000	190.000000	150.000000	1.000000

Tableau	R	Power BI	SAS	Word	PowerPoint	AWS	Python	Bachelor	Master	No_degree	age
2251.000000	2251.000000	2251.000000	2251.000000	2251.000000	2251.000000	2251.000000	2251.000000	2251.000000	2251.000000	2251.000000	2251.000000
0.275433	0.195913	0.079964	0.182586	0.120835	0.086184	0.122168	0.282985	0.471790	0.221235	0.440693	32.066637
0.446832	0.396990	0.271298	0.386412	0.326008	0.280698	0.327552	0.450549	0.499315	0.415171	0.496581	45.457780
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-1.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-1.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	20.000000
1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	38.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	342.000000

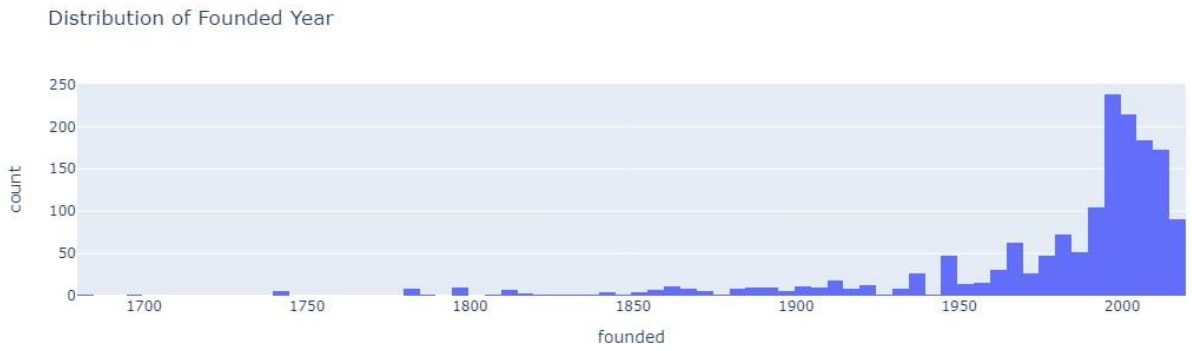
فيما يلي وصف مختصر لقيمنا العددية وسنتناول تفاصيل كل واحدة منها:



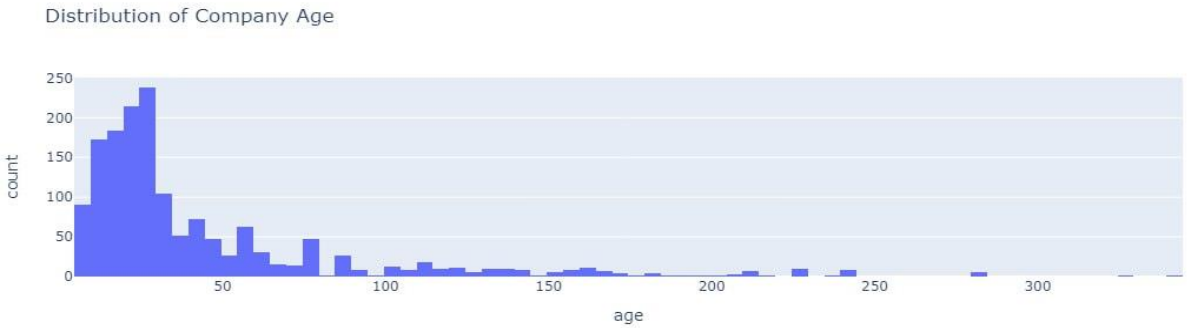


تتضمن المجموعة الأولى من المخططات رسمًا بيانيًا ومخططًا مربعًا لعمود "التقييم". يظهر الرسم البياني أن معظم تقييمات الشركات تتجمع بين 3.0 و4.0، مع ذروة تبلغ حوالي 3.5، مما يشير إلى أن غالبية التقييمات أعلى من المتوسط. يسلط المخطط المربع أدناه الضوء على وجود القيم المتطرفة أقل من 2.0، مع متوسط تصنيف حوالي 3.5.

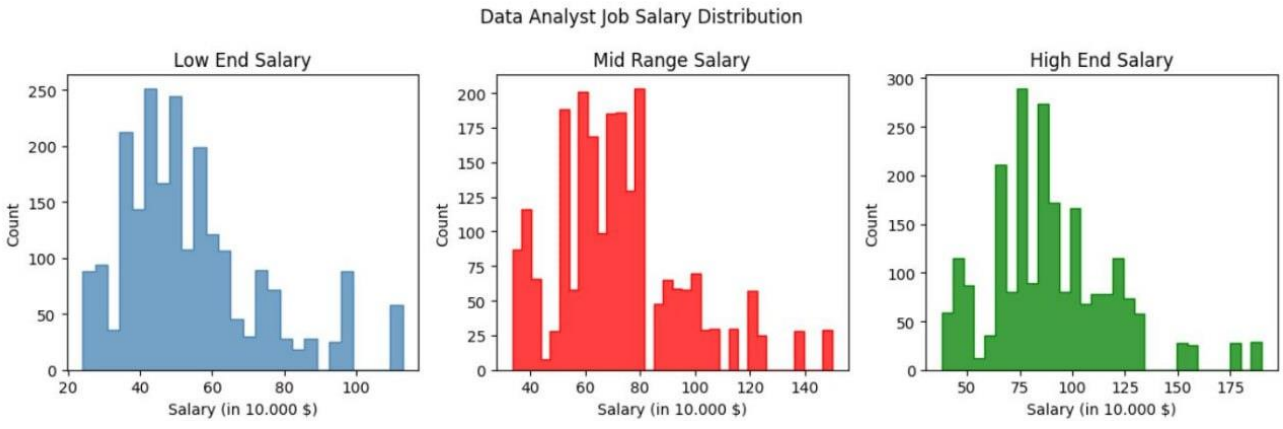
تحتوي المجموعة الثانية من قطع الأراضي على قطعة أرض مربعة أخرى وقطعة كثافة لعمود "التقييم". يؤكد مخطط الصندوق وجود القيم المتطرفة ويظهر توزيعًا بمتوسط يبلغ حوالي 3.5. يوضح مخطط الكثافة هذا التوزيع أيضًا، مع تسليط الضوء على ذروة بارزة عند حوالي 3.5، مع محاذاة المتوسط والوسيط والوضع بشكل وثيق. يشير هذا الاتساق إلى اتجاه تصنيف إيجابي بشكل عام بين الشركات.



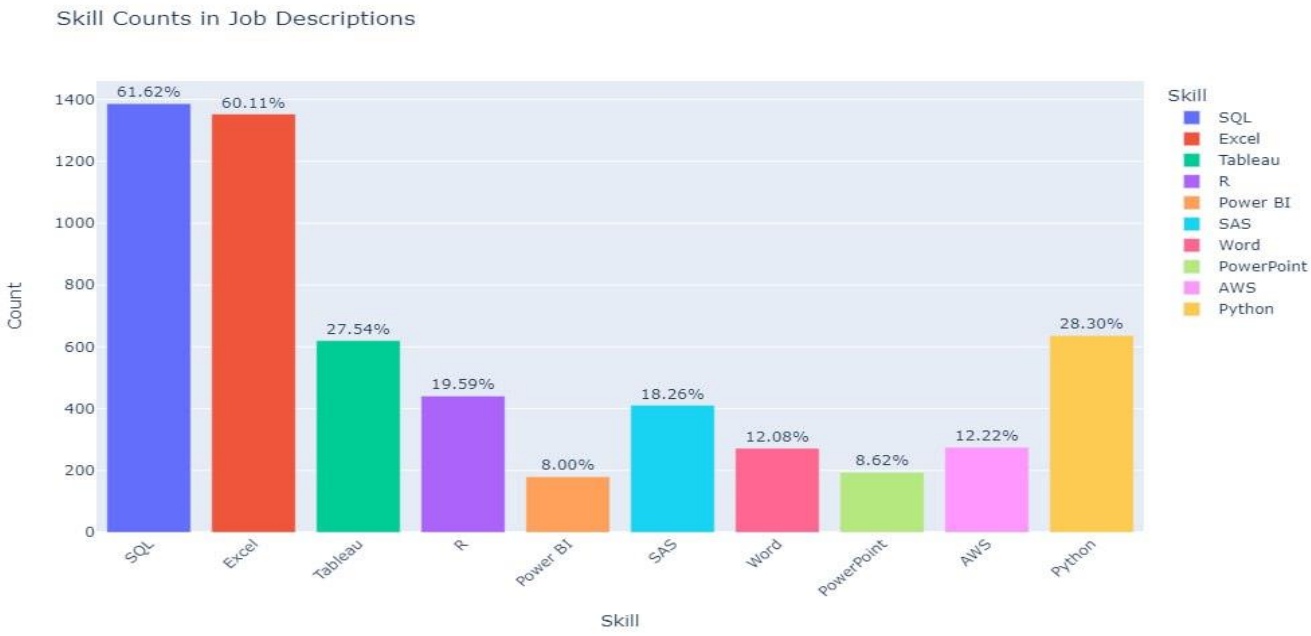
يوضح الرسم البياني توزيع سنوات تأسيس الشركة، مما يشير إلى أن غالبية الشركات في مجموعة البيانات تأسست بعد عام 1950، مع زيادة كبيرة في عدد الشركات التي تأسست حوالي عام 2000. وهناك عدد قليل جدًا من الشركات التي تأسست قبل عام 1900، مما يشير إلى وجود عدد قليل جدًا من الشركات التي تأسست قبل عام 1900. تركيز الشركات الجديدة. قد يعكس الارتفاع المفاجئ في عام 2000 زيادة في الشركات الجديدة خلال الطفرة التكنولوجية.



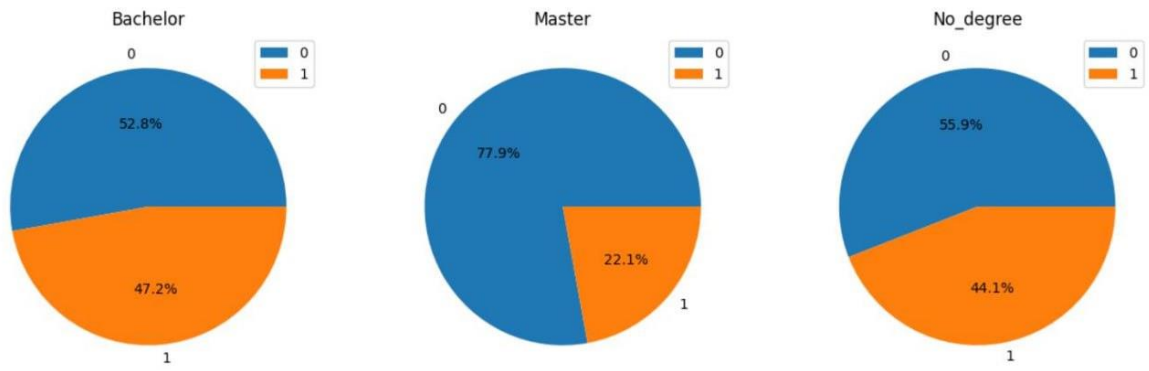
يوضح الرسم البياني توزيع أعمار الشركات، مما يشير إلى أن معظم الشركات حديثة العهد نسبيًا، مع وجود أعلى تركيز لها بين 0 إلى 50 عامًا. يتناقص التكرار بشكل حاد مع تقدم العمر، مع وجود عدد قليل جدًا من الشركات التي يزيد عمرها عن 100 عام. ويشير هذا إلى أن الشركات الأحدث تهيمن على مجموعة البيانات، وأن هناك عددًا أقل بكثير من الشركات القائمة منذ فترة طويلة.



يوضح المخطط توزيع رواتب وظائف محلل البيانات عبر ثلاثة نطاقات: نهاية منخفضة، ومتوسطة المدى، وعالية النهاية. في نطاق الراتب المنخفض، تتجمع معظم الرواتب حول 40.000 دولار - 50.000 دولار، في حين يظهر الراتب متوسط المدى ذروة تبلغ حوالي 70.000 دولار - 80.000 دولار. يشير توزيع الرواتب العالية إلى أن معظم الرواتب تتراوح بين 90,000 دولار - 100,000 دولار، مع وجود حالات أقل للرواتب التي تزيد عن



يوضح الرسم البياني الشريطي مدى تكرار المهارات المحددة المذكورة في الوصف الوظيفي. تعد SQL وExcel من المهارات المطلوبة الأكثر شيوعًا، حيث يظهر كل منها في أكثر من 1000 وصف وظيفي. يتم أيضًا ذكر مهارات أخرى مثل Tableau وR وPython بشكل متكرر، ولكن أقل من SQL وExcel، مما يشير إلى التركيز القوي على هذه المهارات التقنية في إعلانات الوظائف.



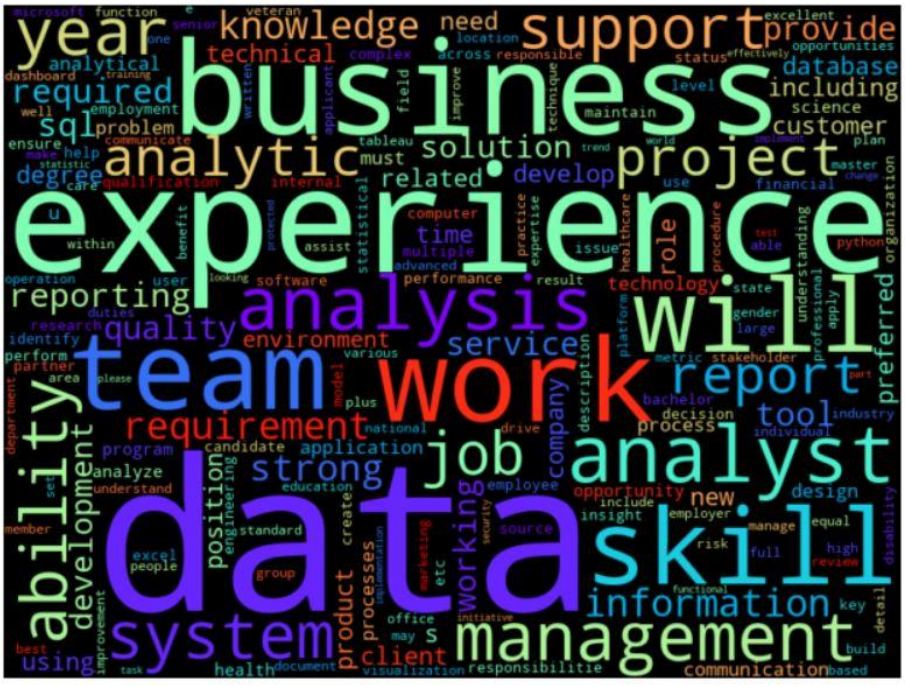
توضح المخططات الدائرية توزيع متطلبات الوظيفة لدرجة البكالوريوس والماجستير وعدم وجود درجة عبر قوائم الوظائف. ما يقرب من 47.2% من قوائم الوظائف تتطلب درجة البكالوريوس، في حين أن 22.1% تتطلب درجة الماجستير. ومن الجدير بالذكر أن 44.1% من قوائم الوظائف لا تتطلب أي درجة علمية، مما يشير إلى أن جزءًا كبيرًا من فرص العمل يمكن الوصول إليها دون الحصول على أوراق اعتماد رسمية من التعليم العالي.

B. القيم الفئوية:

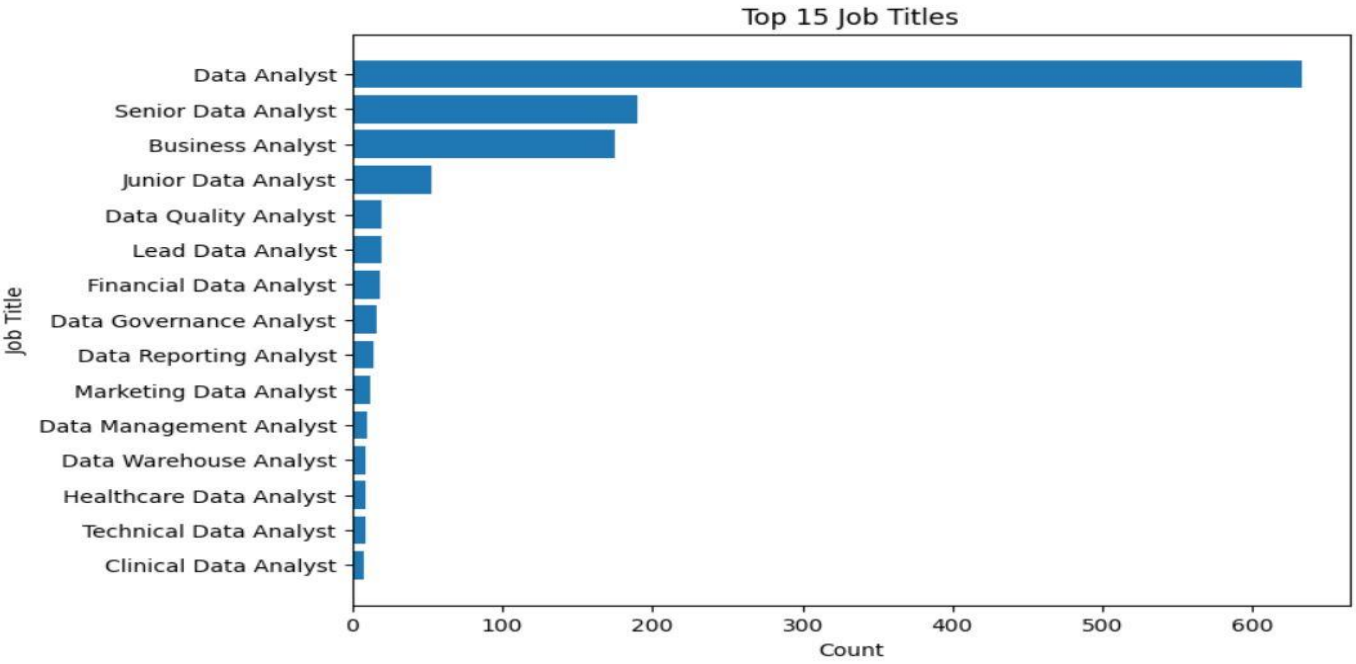
	job_title	salary_estimate	job_description	company_name	location	headquarters	size
count	2251	2251	2251	2251	2251	2251	2251
unique	833	89	2251	1500	253	496	8
top	Data Analyst	42 - 76	Are you eager to roll up your sleeves and harn...	Staffigo Technical Services, LLC	New York, NY	New York, NY	51 to 200 employees
freq	633	57	1	58	310	227	421

type_of_ownership	sector	revenue	average_salary_bin	job_city	job_state	hq_city	hq_state	age_bin
2251	2251	2251	2251	2251	2251	2251	2251	2251
14	25	13	3	249	19	479	68	4
Company - Private	Information Technology	Unknown / Non-Applicable	high	New York	CA	New York	CA	old
1273	570	776	1074	310	626	227	512	664

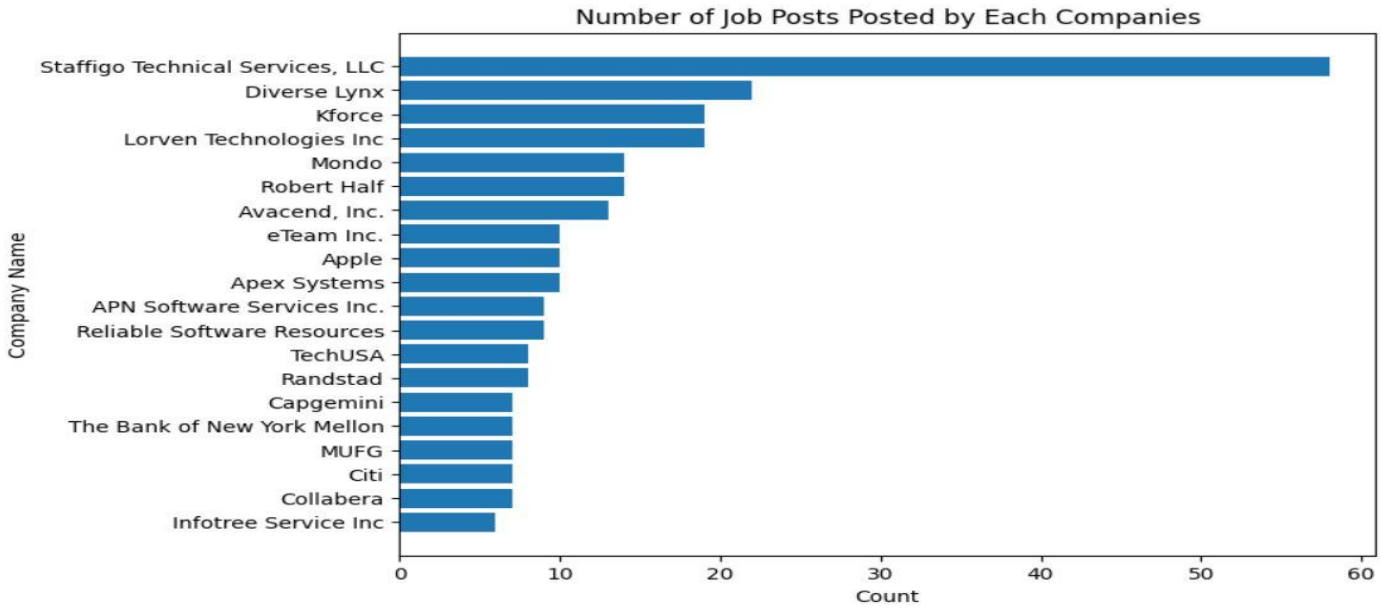
فيما يلي وصف موجز لقيمنا الفئوية وسنتناول تفاصيل كل واحدة منها:



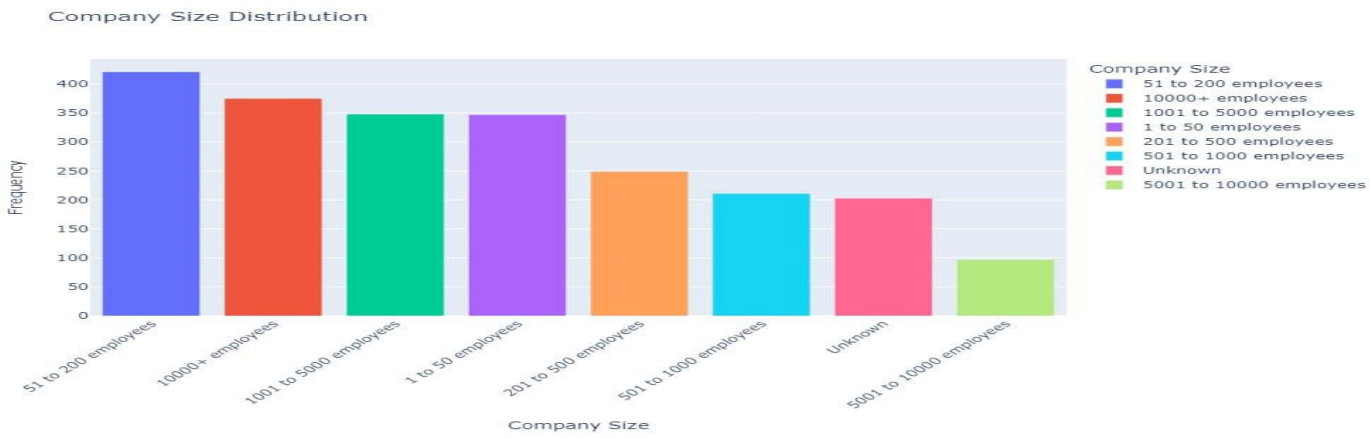
تسلط سحابة الكلمات الضوء على المصطلحات الأكثر تكرارًا في job_description، حيث تشير الكلمات الأكبر حجمًا إلى تكرار أعلى. تشير المصطلحات الأساسية مثل "البيانات"، و"الخبرة"، و"الأعمال"، و"المهارات"، و"الفريق" إلى التركيز القوي على الخبرة المتعلقة بالبيانات، والخبرة العملية، والعمل الجماعي في متطلبات الوظيفة. يشير هذا التصور إلى أن التوصيف الوظيفي يعطي الأولوية للمرشحين ذوي الخبرة والمهارات ذات الصلة في تحليل البيانات والعمليات التجارية.



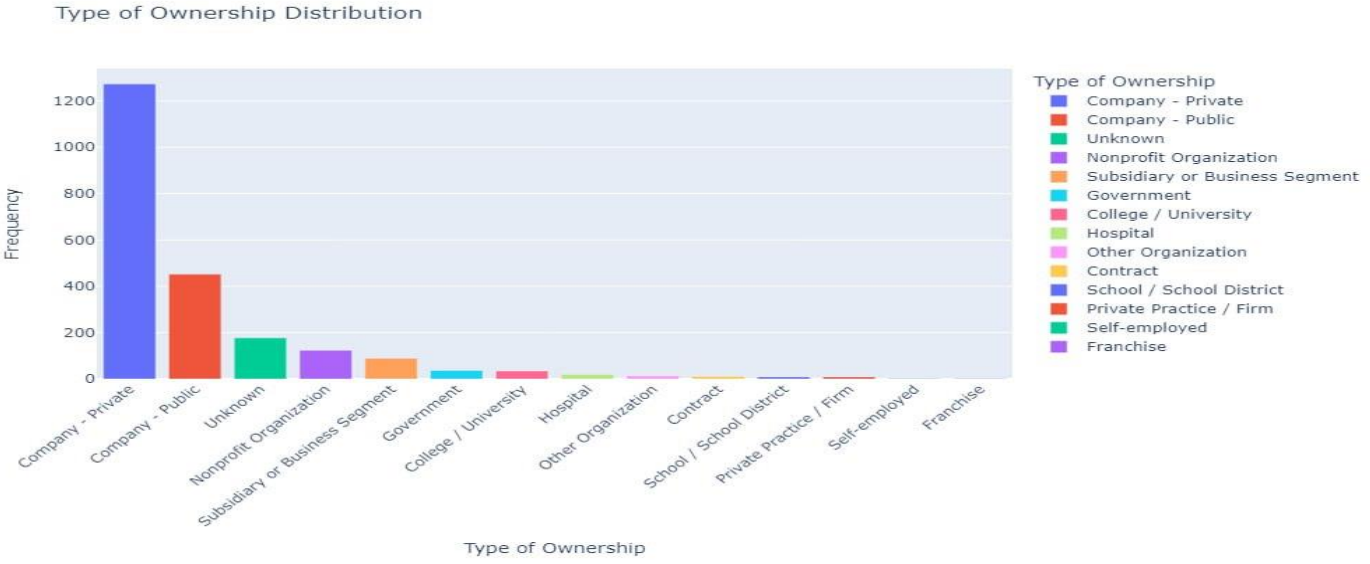
هنا نرى أن Data Analyst هو المسمى الوظيفي الأكثر



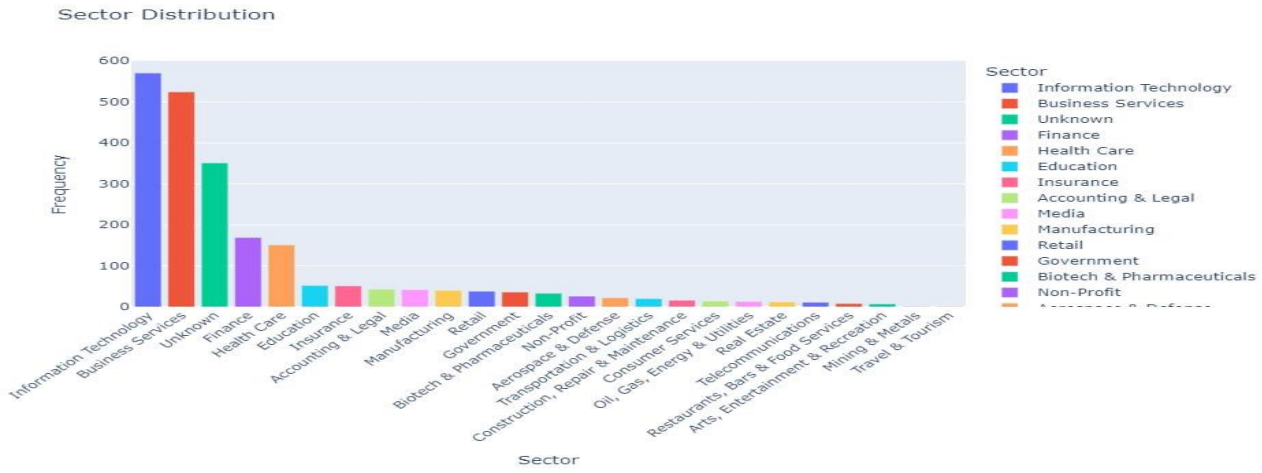
تنشر شركة Staffigo Technical Services, LLC معظم منشورات وظائف محلل البيانات في البيانات.



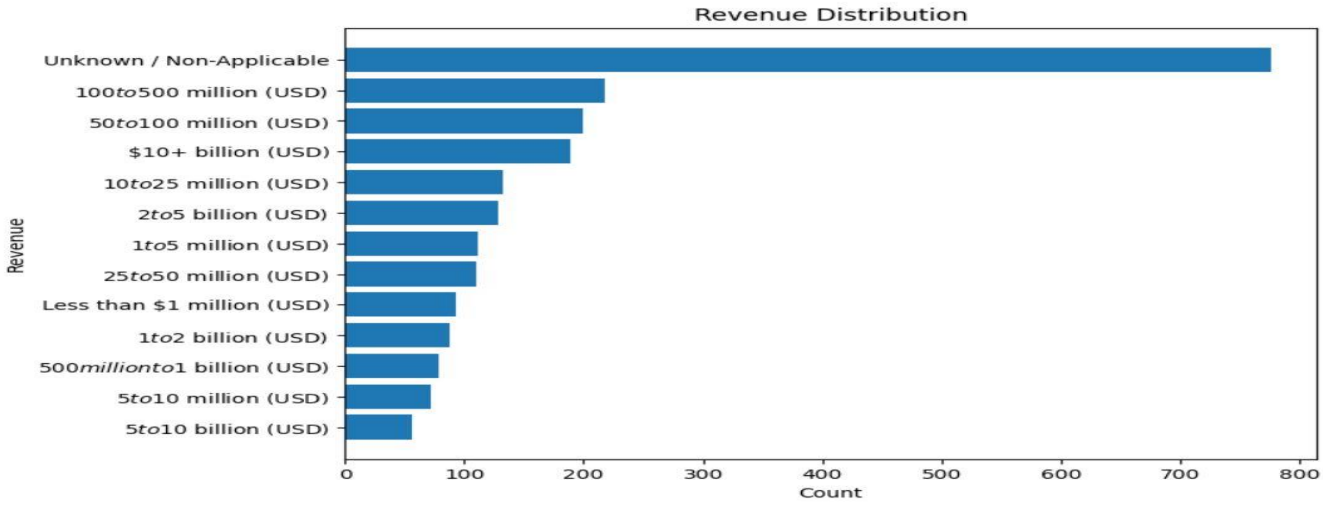
يوضح الرسم البياني الشريطي توزيع أحجام الشركات، حيث تضم غالبية الشركات ما بين 51 إلى 200 موظف، تليها مباشرة الشركات التي تضم أكثر من 10000 موظف. الشركات متوسطة الحجم، مثل تلك التي تضم 1001 إلى 5000 موظف ومن 1 إلى 50 موظفًا، تتمتع أيضًا بتمثيل كبير. والفئة التي تضم أقل عدد من الشركات هي من 5001 إلى 10000 موظف، مما يشير إلى وجود أقل للشركات ضمن هذا النطاق الحجمي.



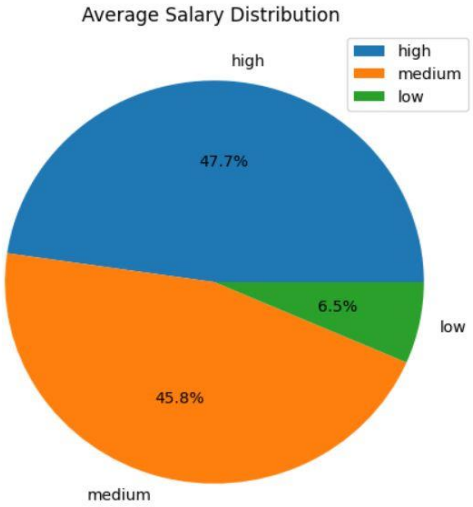
ويبين الرسم البياني توزيع أنواع الملكية المختلفة بين الشركات، وأغلبها شركات خاصة، والتي تمثل أكثر من 1200 مدخل. الشركات العامة هي ثاني أكثر الشركات شيوعًا، يليها عدد ملحوظ من الإدخالات المصنفة على أنها "غير معروفة" و"منظمة غير ربحية". أنواع الملكية الأخرى، مثل الشركات التابعة، والمؤسسات الحكومية والتعليمية، لديها عدد أقل بكثير من الإدخالات.



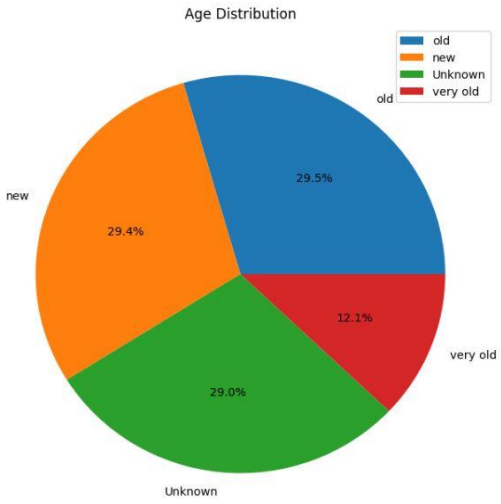
يعرض الرسم البياني الشريطي توزيع القطاعات بين الشركات، مع كون تكنولوجيا المعلومات وخدمات الأعمال هي الأكثر انتشارًا، حيث يحتوي كل منها على أكثر من 500 إدخال. تحتوي الفئة "غير معروف" أيضًا على عدد كبير من الإدخالات، مما يشير إلى بيانات غير كاملة للعديد من الشركات. ويتم تمثيل قطاعات أخرى مثل المالية والرعاية الصحية والتعليم بدرجة أقل، في حين تظهر العديد من القطاعات الأصغر ترددات أقل بكثير.



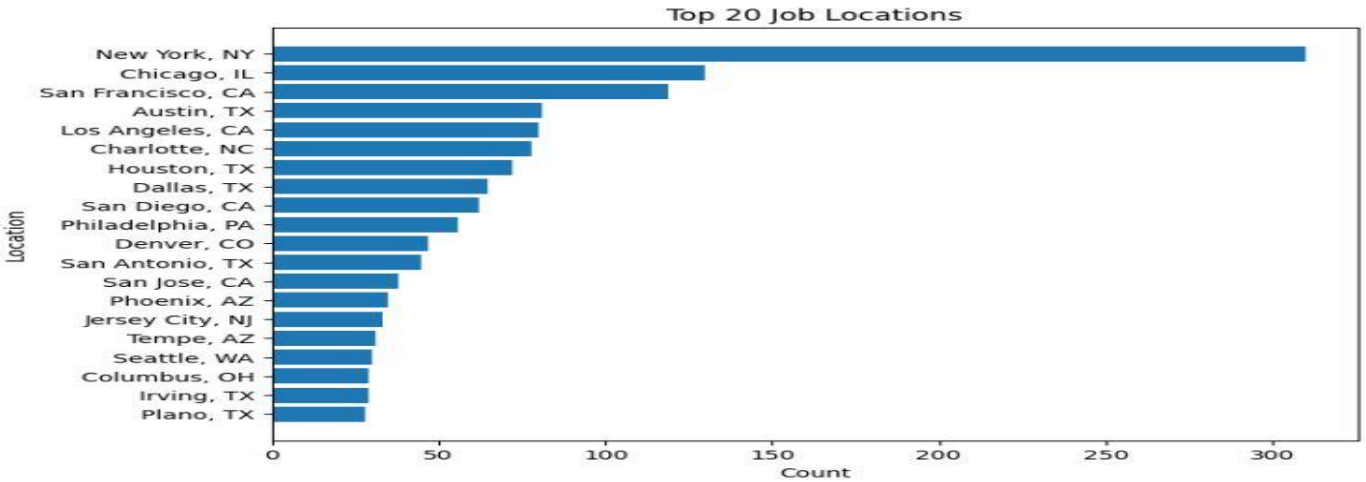
ويوضح الرسم البياني الشريطي توزيع إيرادات الشركات، حيث يظهر أن الأغلبية لديها إيرادات غير معروفة أو غير قابلة للتطبيق، تليها الشركات التي تتراوح أرباحها بين 100 إلى 500 مليون دولار أمريكي، ومن 50 إلى 100 مليون دولار أمريكي. الشركات ذات الإيرادات المرتفعة (+10 مليار دولار أمريكي) وتتراوح الإيرادات المتوسطة مثل 10 إلى 25 مليون دولار أمريكي لديها أيضًا تمثيلات كبيرة. الفئات ذات الإيرادات المنخفضة، مثل تلك التي تكسب أقل من مليون دولار أمريكي، لديها عدد أقل من الإدخالات، مما يشير إلى وجود أقل للشركات الصغيرة في مجموعة البيانات.



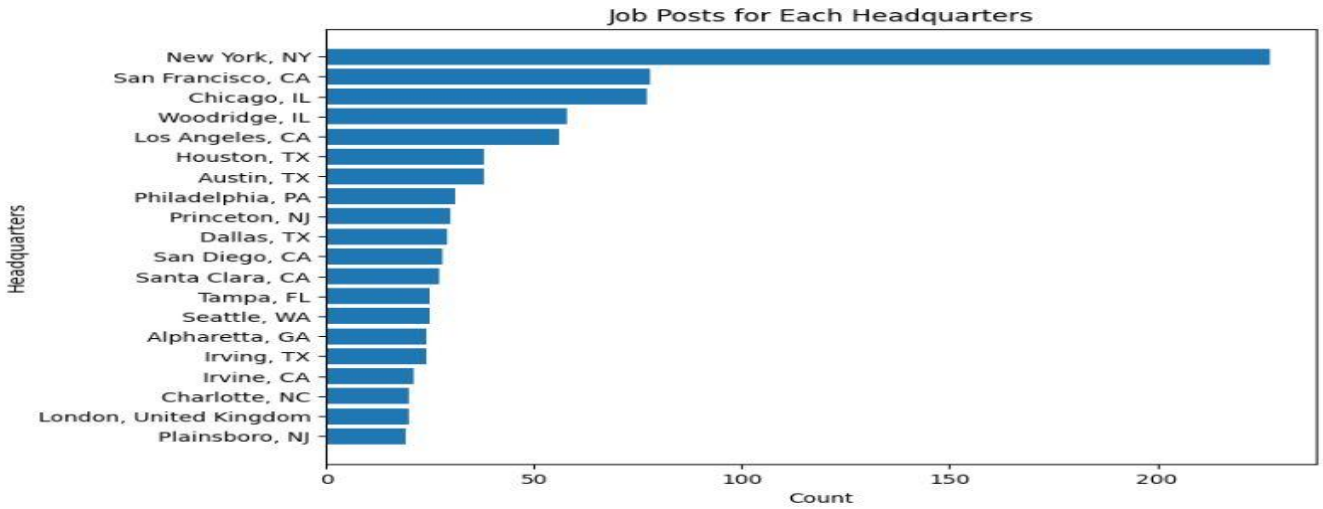
يوضح الرسم البياني الدائري توزيع متوسط الرواتب، موضحًا أن ما يقرب من نصف (47.7٪) من الإدخالات يقع ضمن فئة الرواتب المرتفعة. أما فئة الراتب المتوسط فهي أقل انتشارًا بقليل، حيث تمثل 45.8% من المشاركات. فئة الرواتب المنخفضة هي الأقل شيوعًا، حيث تمثل 6.5% فقط من الإدخالات، مما يشير إلى أن معظم الوظائف تقدم رواتب متوسطة إلى عالية في المتوسط.



يوضح الرسم البياني الدائري توزيع أعمار الشركات، حيث تشكل كل من الشركات "القديمة" و"الجديدة" ما يقرب من 30% من مجموعة البيانات. وتمثل الفئة "غير معروف" أيضًا نسبة كبيرة تبلغ 29%، مما يشير إلى عدم اكتمال البيانات العمرية للعديد من الشركات. وتشكل الشركات "القديمة جداً" 12.1%، مما يشير إلى أن نسبة أقل من الشركات تم تأسيسها منذ فترة طويلة.



نيويورك لديها حتى الآن أكبر عدد من عروض العمل



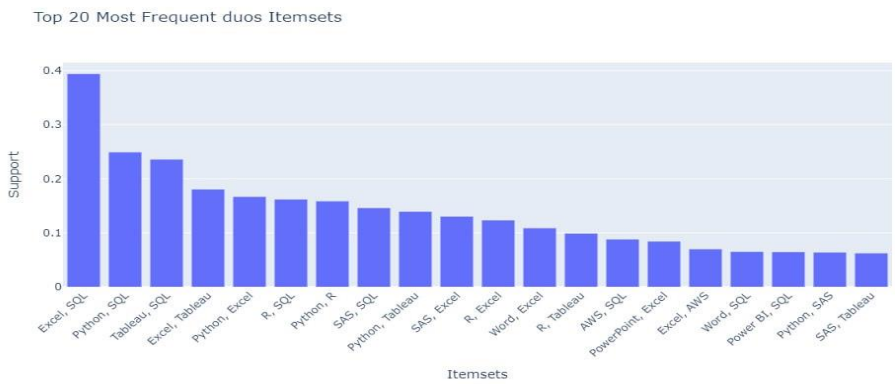
نيويورك لديها أكبر عدد من المقرات

---

(2) والان سنقوم بالتحليل ثنائي المتغير (bivariant):

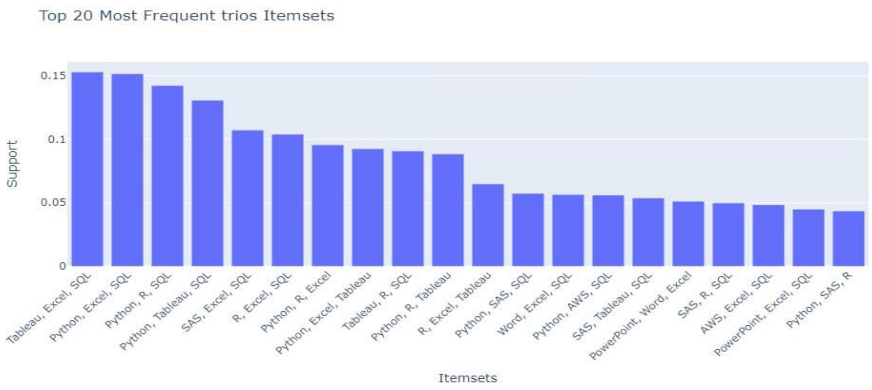
أولاً نختار أعمدة المهارات ونحولها إلى منطقية ثم نطبق خوارزمية Apriori للعثور على مجموعات العناصر المتكررة (مجموعات مشتركة من المهارات) مع حد أدنى من الدعم يبلغ 0.2%. ثم يتم إضافة عمود طول جديد إلى النتيجة للإشارة إلى عدد العناصر في كل عنصر متكرر.

	support	itemsets	length
10	0.394047	(Excel, SQL)	2
18	0.249223	(Python, SQL)	2
11	0.235895	(Tableau, SQL)	2
19	0.180809	(Excel, Tableau)	2
26	0.167037	(Python, Excel)	2

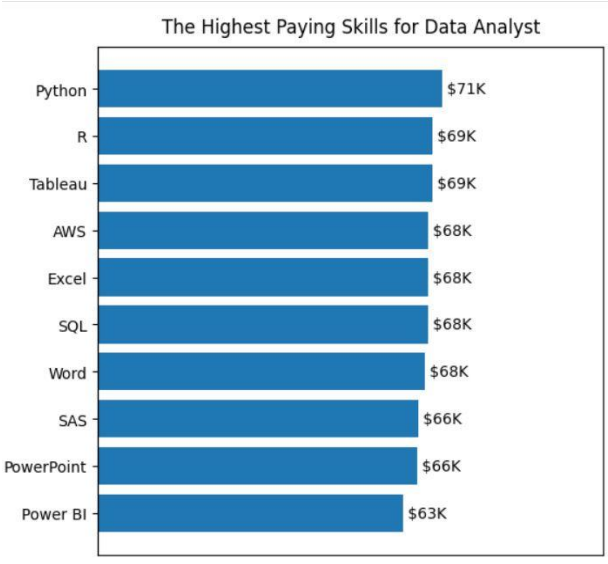


يوضح المخطط الشريطي أهم 20 مجموعة من المهارات المطلوبة في إعلانات الوظائف، مع كون "Excel وSQL" هي التركيبة الأكثر شيوعًا، والتي تظهر في ما يقرب من 40% من الإعلانات. تتضمن المجموعات السائدة الأخرى "Python وSQL"، و"Tableau and SQL"، و"Excel وTableau"، ولكل منها قيم دعم كبيرة. تسلط الحبكة الضوء على أهمية مجموعات مهارات محددة، خاصة تلك التي تتضمن Excel وSQL وPython، في سوق العمل.

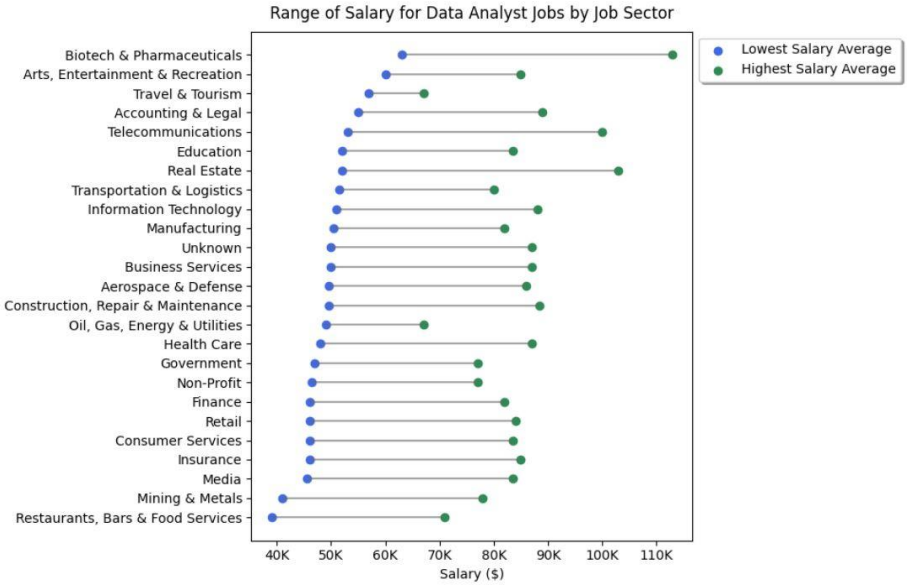
	support	itemsets	length
55	0.152821	(Tableau, Excel, SQL)	3
62	0.151488	(Python, Excel, SQL)	3
75	0.142159	(Python, R, SQL)	3
69	0.130609	(Python, Tableau, SQL)	3
58	0.107064	(SAS, Excel, SQL)	3



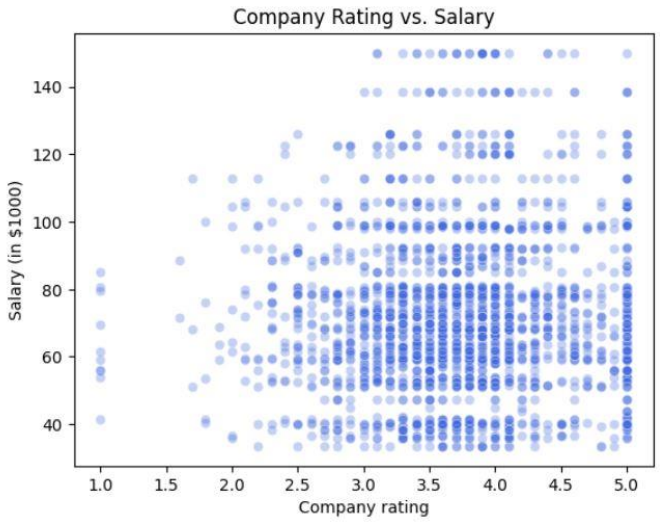
يُظهر المخطط الشريطي أهم 20 عنصرًا من المهارات المطلوبة في إعلانات الوظائف، مع كون "Excel وTableau وSQL" هي التركيبة الأكثر شيوعًا، حيث تظهر في حوالي 15% من الإعلانات. تتضمن المجموعات البارزة الأخرى "Excel وPython وSQL" و"R وPython وSQL" و"Excel وSAS وSQL". تسلط القصة الضوء على الطلب المرتفع على مجموعات محددة من المهارات المتعلقة بالبيانات، خاصة تلك التي تتضمن Excel وSQL وPython وTableau وR.



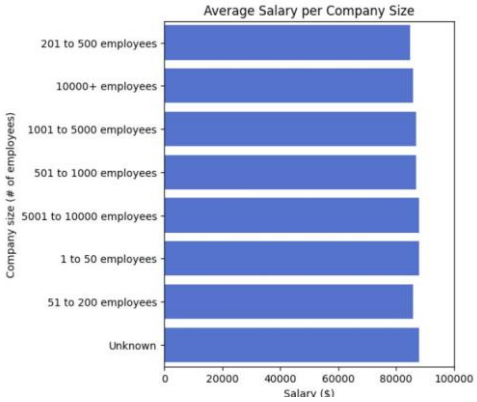
يُظهر الرسم البياني الشريطي المهارات الأعلى أجرًا لمحللي البيانات، حيث تتقدم لغة Python بمتوسط راتب قدره 71 ألفًا. تتبع مهارات مثل R وTableau بشكل وثيق 69 ألفًا، في حين أن الأدوات الأساسية الأخرى مثل AWS وExcel وSQL وWord وSAS تقدم متوسط رواتب يبلغ حوالي 68 ألفًا. على الرغم من أن PowerPoint وPower BI لا يزالان قيمين، إلا أن متوسط الرواتب لديهما أقل قليلاً عند 66 ألفًا و63 ألفًا على التوالي.



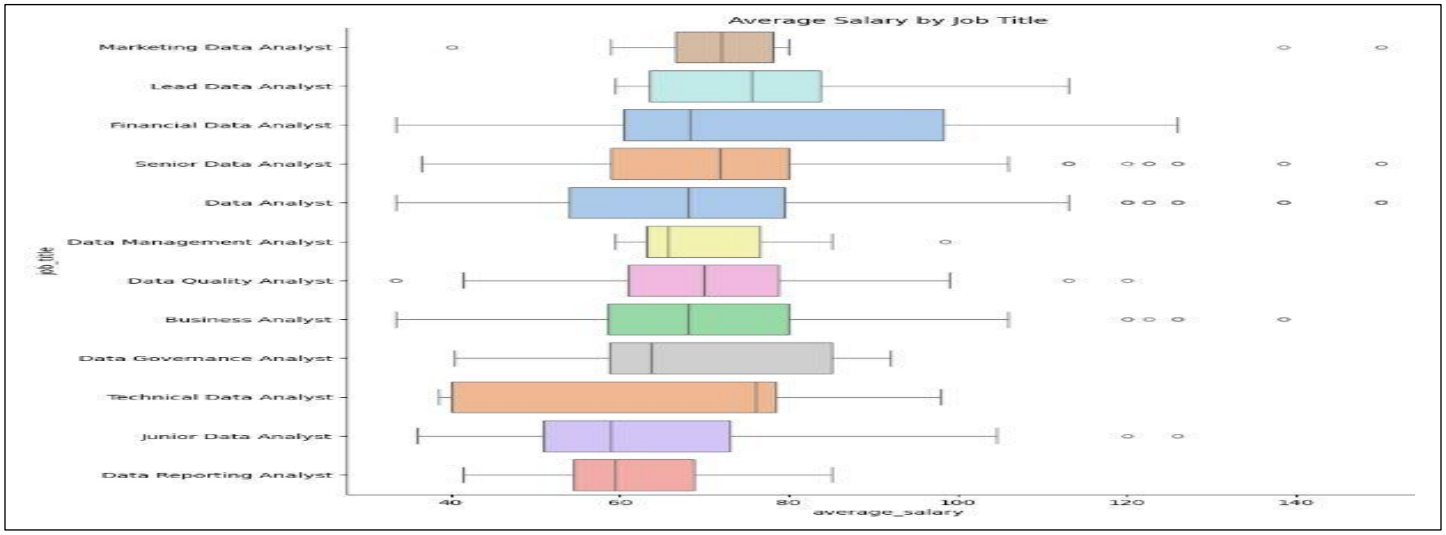
يعرض مخطط النقاط نطاق الراتب لوظائف محلل البيانات عبر قطاعات العمل المختلفة، حيث تمثل النقاط الزرقاء متوسط الراتب الأدنى والنقاط الخضراء تمثل أعلى متوسط الراتب. تُظهر قطاعات مثل التكنولوجيا الحيوية والمستحضرات الصيدلانية والفنون والترفيه والترفيه والسفر والسياحة أكبر نطاقات للرواتب تصل إلى حوالي 110 ألف دولار. كما تقدم قطاعات أخرى، مثل تكنولوجيا المعلومات والمالية، متوسطات رواتب عالية، مما يسلط الضوء على التنوع في التعويضات عبر مختلف الصناعات.



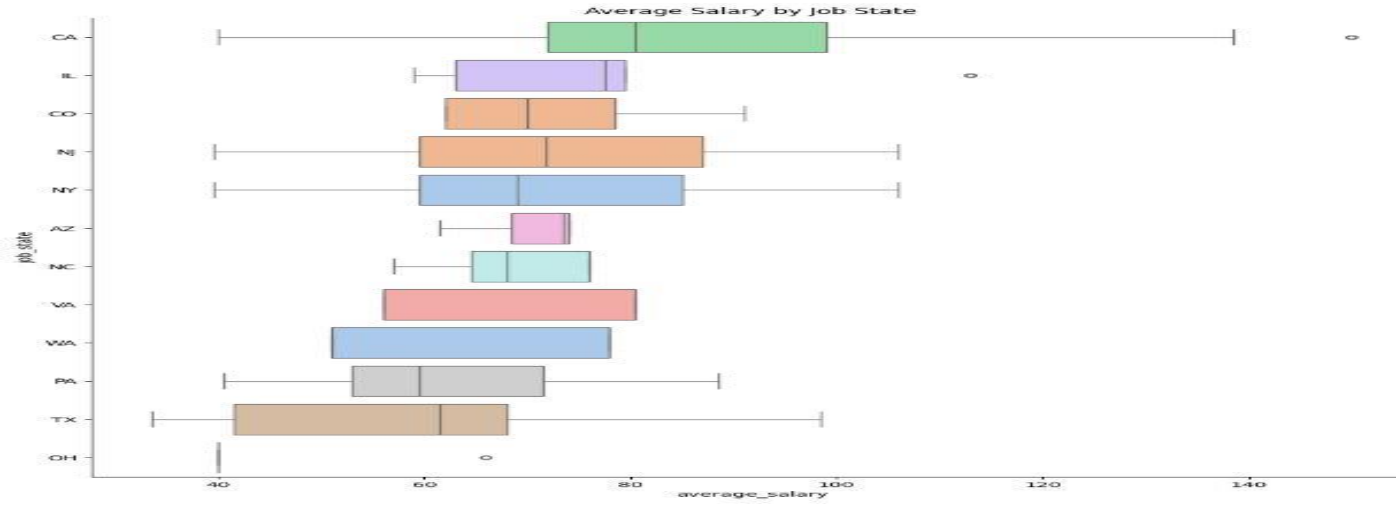
يوضح المخطط المبعثر العلاقة بين تصنيفات الشركة والرواتب، مما يوضح توزيعًا واسعًا للرواتب عبر جميع مستويات التصنيف. في حين أن معظم الرواتب تتراوح بين 50 ألفًا و100 ألف بغض النظر عن التصنيف، إلا أن هناك قيمًا متطرفة أعلى من 120 ألفًا، خاصة في الطرف الأعلى من مقياس التصنيف. تشير المؤامرة إلى أنه في حين أن الشركات ذات التصنيف الأعلى تميل إلى تقديم مجموعة واسعة من الرواتب، فإن الرواتب المرتفعة لا ترتبط حصريًا بأعلى تصنيفات الشركة.



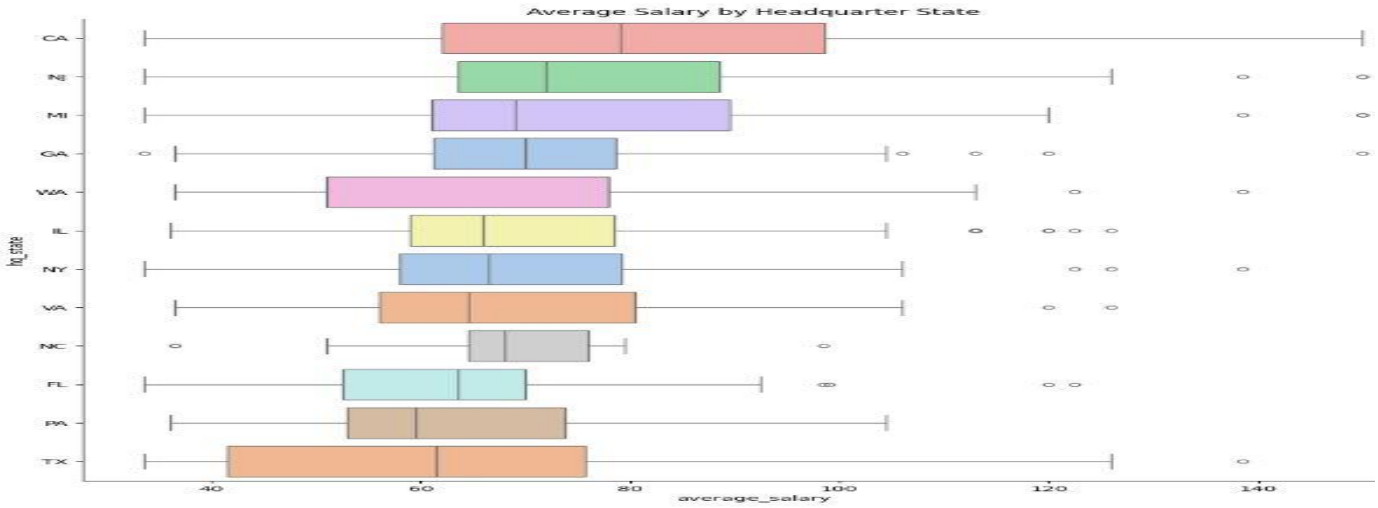
حجم الشركة لا يعني الكثير من حيث الراتب لوظيفة محلل البيانات.



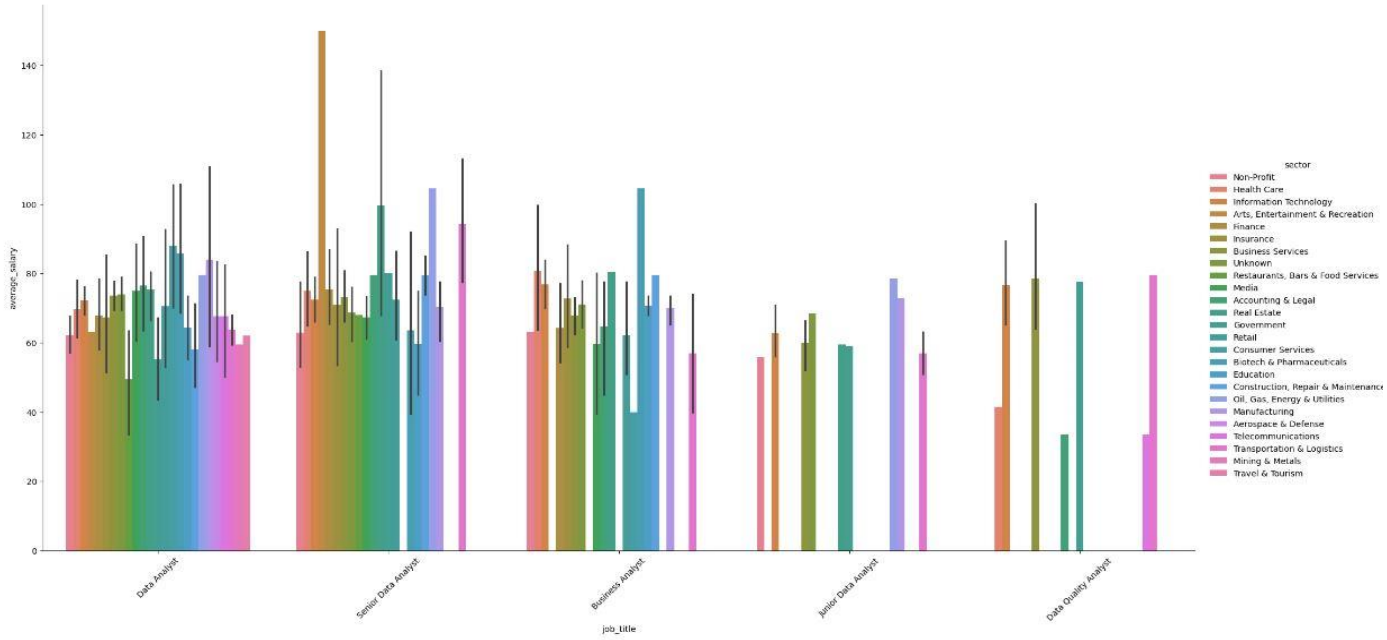
يوضح المخطط المربع أن أدوار محلل البيانات الأول، ومحلل البيانات الرئيسي، ومحلل البيانات الفنية تتمتع بأعلى الرواتب المتوسطة والتباين الكبير. معظم المسميات الوظيفية، بما في ذلك محلل البيانات ومحلل البيانات المالية، لديها متوسط رواتب يتراوح بين 60 ألف دولار إلى 80 ألف دولار. الأدوار ذات الأجور المنخفضة مثل محلل تقارير البيانات ومحلل البيانات المبتدئ لها رواتب متوسطة أقل، مما يشير إلى نطاق رواتب متنوع عبر وظائف محلل البيانات المختلفة



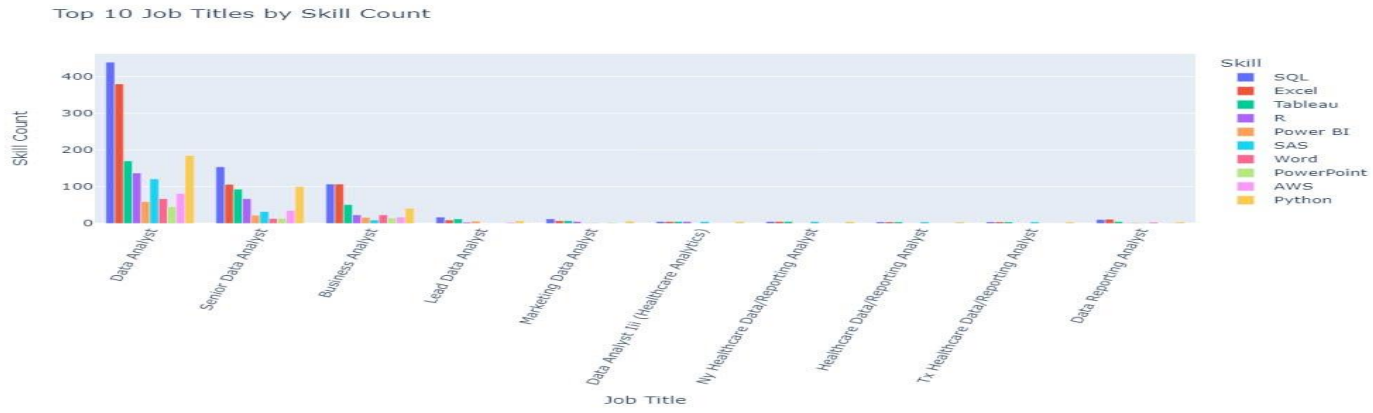
يُظهر المخطط المربع متوسط الرواتب حسب الولاية الوظيفية، مما يشير إلى أن كاليفورنيا (كاليفورنيا) ونيوجيرسي (نيوجيرسي) لديهما أعلى متوسط رواتب وتباين واسع. كما تقدم ولايات مثل نيويورك (نيويورك) وفيرجينيا (فيرجينيا) وواشنطن (واشنطن) رواتب متوسطة عالية. بعض الولايات، مثل أريزونا (أريزونا) ونورث كارولينا (نورث كارولاينا)، لديها متوسط رواتب أقل ونطاقات رواتب أصغر. هناك قيم متطرفة ملحوظة في إلينوي (إلينوي) وكاليفورنيا (كاليفورنيا)، مما يشير إلى وجود عدد قليل من الرواتب المرتفعة بشكل استثنائي.



يُظهر المخطط المربع متوسط الرواتب حسب الولاية الرئيسية، حيث تقدم كاليفورنيا (كاليفورنيا) ونيوجيرسي (نيو جيرسي) أعلى متوسط رواتب ونطاقات كبيرة للرواتب. تتمتع ولايات مثل ميشيغان (MI)، وواشنطن (WA)، ونيويورك (NY) أيضًا برواتب متوسطة مرتفعة نسبيًا وتقلب ملحوظ. تتمتع كارولينا الشمالية (نورث كارولاينا) وفلوريدا (فلوريدا) برواتب متوسطة أقل ونطاقات رواتب أضيق. توجد القيم المتطرفة في العديد من الولايات، وخاصة إلينوي (إلينوي) وكاليفورنيا (كاليفورنيا)، مما يشير إلى بعض الرواتب المرتفعة بشكل استثنائي داخل تلك الولايات.

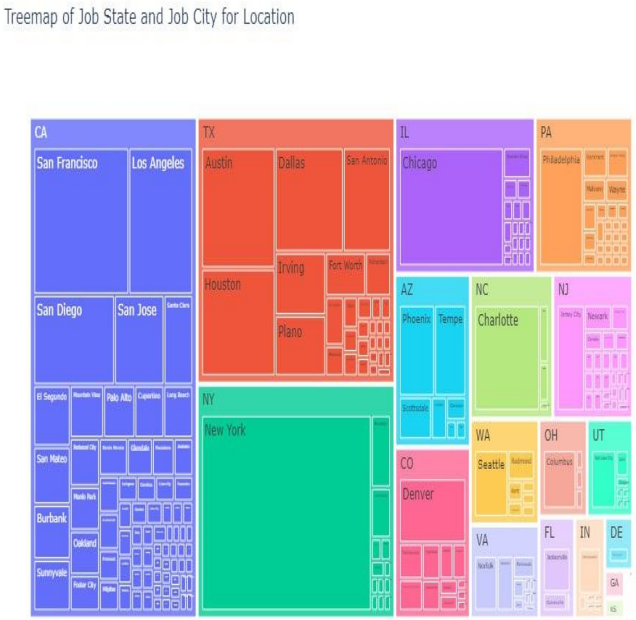


يُظهر المخطط الشريطي متوسط الراتب لمختلف المسميات الوظيفية، مصنفة حسب القطاعات المختلفة. ويكشف عن تباين كبير في الرواتب ضمن كل مسمى وظيفي، متأثرًا بالقطاع. على سبيل المثال، تظهر أدوار "محلل بيانات كبير" نطاقات رواتب أعلى مقارنة بأدوار "محلل بيانات مبتدئ"، مع ملاحظة أعلى الرواتب في قطاعات مثل تكنولوجيا المعلومات والمالية.



يُظهر المخطط الشريطي أعلى 10 مسميات وظيفية حسب عدد المهارات، مع حصول "محلل البيانات" على أعلى عدد إجمالي من المهارات عبر جميع المهارات المدرجة. والجدير بالذكر أن SQL وExcel وPython هي المهارات الأكثر ذكرًا في أدوار محلل البيانات. تتطلب المسميات الوظيفية الأخرى مثل " Senior Data Analyst" و"Business Analyst" هذه المهارات أيضًا، ولكن بدرجة أقل. يسلط المخطط الضوء على الطلب على مجموعة واسعة من المهارات عبر المسميات الوظيفية المختلفة، مع التركيز على أهمية SQL وExcel وPython في الأدوار المتعلقة بالبيانات.

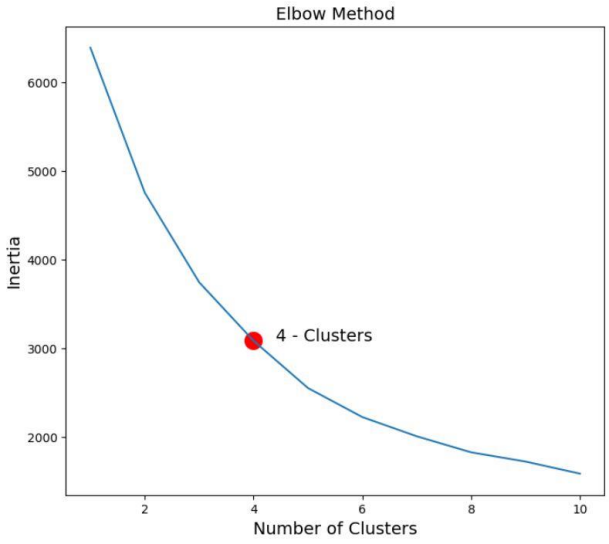
يعرض المخطط الهيكلي قوائم الوظائف حسب الولاية والمدينة، مما يوضح أن ولاية كاليفورنيا (كاليفورنيا) لديها أعلى تركيز لفرص العمل، خاصة في سان فرانسيسكو ولوس أنجلوس. تكساس (تكساس) ونيويورك (نيويورك)، مع المدن الكبرى مثل أوستن ودالاس ومدينة نيويورك، لديها أيضًا قوائم وظائف مهمة. تتمتع ولايات أخرى مثل إلينوي (IL) وبنسلفانيا (PA) وأريزونا (AZ) بمجموعات ملحوظة ولكنها أصغر من فرص العمل.



يُظهر المخطط الهيكلي توزيع المقر الرئيسي للشركة حسب الولاية والمدينة، مع وجود أهم المجموعات في كاليفورنيا (كاليفورنيا) ونيويورك (نيويورك)، خاصة في سان فرانسيسكو ولوس أنجلوس ومدينة نيويورك. ولايات أخرى مثل إلينوي (IL)، وتكساس (TX)، ونيوجيرسي (NJ) لديها أيضًا تركيزات ملحوظة من مقرات الشركة، مع مدن رئيسية مثل شيكاغو وأوستن ودالاس وبرينستون. ويمكن رؤية مجموعات أصغر عبر مختلف الولايات والمواقع الدولية، مما يشير إلى انتشار متنوع لمقر الشركة.

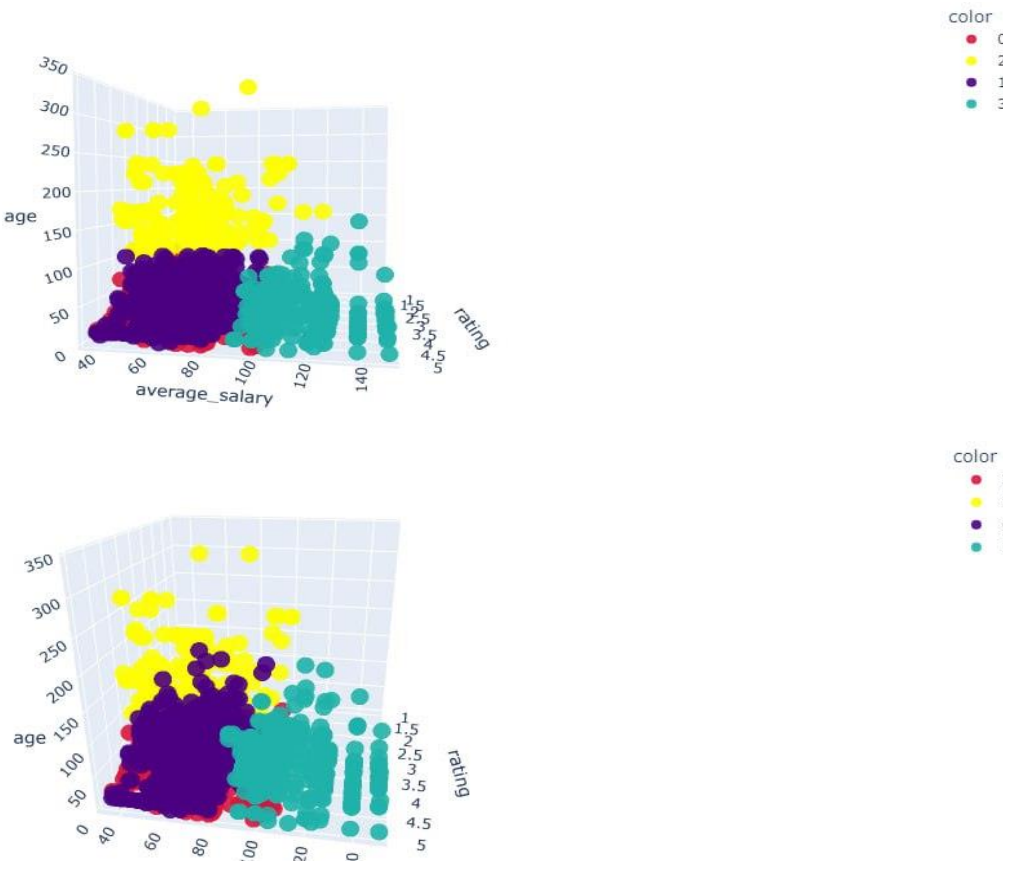
يعرض المخطط المبعثر متوسط الراتب حسب اسم الشركة، حيث تمثل كل نقطة متوسط راتب الشركة لنشر الوظيفة. يشير لون كل نقطة إلى درجة تقييم الشركة، مع تدرج اللون من اللون الأرجواني (التقييمات المنخفضة) إلى الأصفر (التقييمات العالية). تعرض شركات مثل Telligen وChinese Community Health Plan مجموعة واسعة من الرواتب، بتقييمات متفاوتة. يسلط المخطط الضوء على العلاقة بين تقييمات الشركة ومتوسط الرواتب، مما يوضح أن الرواتب المرتفعة تتوزع عبر درجات تصنيف مختلفة.

: cluster analysis (3)

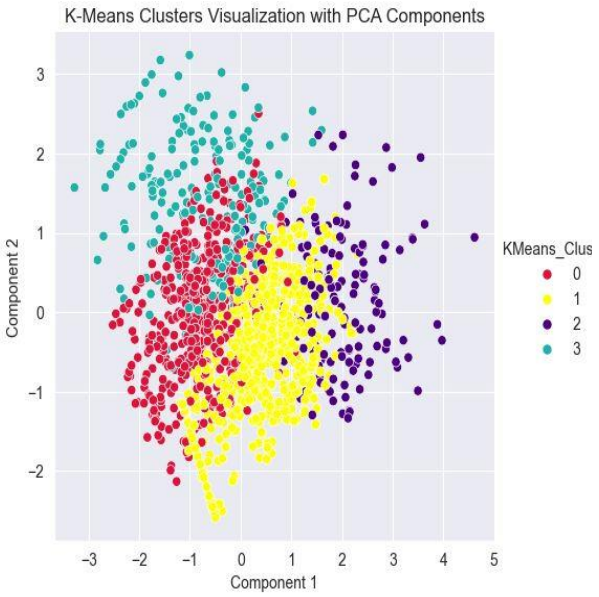


: Elbow Method Plot

يُظهر المخطط طريقة الكوع لتحديد العدد الأمثل للمجموعات لتجميع الوسائل K. يمثل المحور x عدد المجموعات، ويوضح المحور y القصور الذاتي، وهو مقياس لمجموع المربعات داخل المجموعة. تشير النقطة الحمراء عند 4 مجموعات إلى العدد الأمثل للمجموعات التي يحدث فيها الكوع، مما يشير إلى أفضل مفاضلة بين القصور الذاتي وعدد المجموعات.

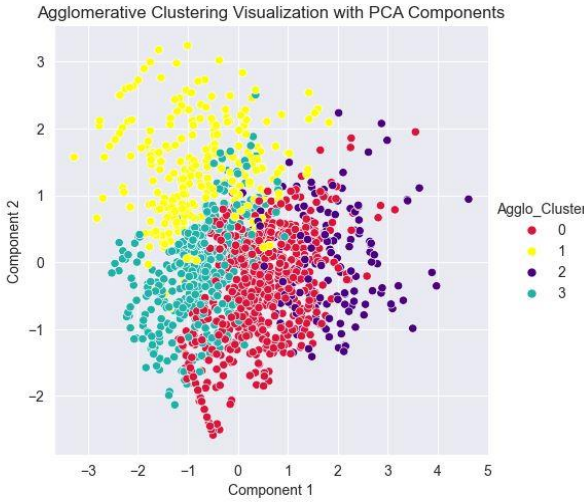


مخطط مبعثر ثلاثي الأبعاد لمجموعات K-Means (3D Scatter Plot) ومخطط مبعثر ثلاثي الأبعاد للمجموعات التكتلية تمثل الألوان المختلفة مجموعات مختلفة، مع أربع مجموعات متميزة توضح نتائج المجموعات الهرمية.



تصور مكونات PCA باستخدام مجموعات K-Means:

يظهر مخطط المبعثر ثنائي الأبعاد (2D Scatter Plot) مكونات PC بعد تقليل الأبعاد لتصور مجموعات K-means. يمثل المحور X المكون الرئيسي الأول، ويمثل المحور Y المكون الرئيسي الثاني. تشير الألوان المختلفة إلى المجموعات الأربع التي تم تحديدها بواسطة وسائل K، مما يوفر رؤية واضحة لفصل المجموعة بأبعاد مخفضة.



تصور مكونات PCA مع المجموعات التكتلية:

يعرض مخطط المبعثر ثنائي الأبعاد (2D Scatter Plot) مكونات PCA لتصور نتائج التجميع التجميعي. المحور x هو المكون الرئيسي الأول، والمحور y هو المكون الرئيسي الثاني. يستخدم المخطط ألوانًا مختلفة للإشارة إلى المجموعات الأربع التي شكلتها التكتلات التكتلية، مما يوضح كيفية توزيع المجموعات في مساحة الأبعاد المخفضة.

	Component 1	Component 2	KMeans_Cluster	Agglo_Cluster
0	0.462109	0.075231	0	3
1	1.778880	-0.626398	2	2
2	0.696282	-0.593179	1	0
3	0.112661	-1.227955	1	0
4	0.192522	-1.066808	1	0
5	0.127818	-1.081864	1	0
6	-0.130067	-1.498281	1	0
7	1.592346	-0.598520	2	2
8	0.621478	-0.457941	2	2
9	-0.285357	-1.534415	1	0

المكون 1 والمكون 2: توضح هذه الأعمدة كيفية تمثيل نقاط البيانات في مساحة ثنائية الأبعاد جديدة بعد PCA. يتوافق كل صف مع إحداثيات نقطة البيانات في هذه المساحة.

KMeans_Cluster: يشير إلى المجموعة التي تنتمي إليها كل نقطة بيانات وفقًا لخوارزمية تجميع الوسائل K.

Agglo_Cluster: يشير إلى المجموعة التي تنتمي إليها كل نقطة بيانات وفقًا لخوارزمية التجميع التجميعي.

---

5. Modeling:

- استخدمنا هنا خط أنابيب شامل للتعلم الآلي لمعالجة مهمة الانحدار لدينا. بدأت العملية بالمعالجة المسبقة للبيانات، والتي تضمنت قياس الميزات وترميزها. تم توحيد الميزات الرقمية باستخدام StandardScaler، في حين تم تحويل الميزات الفئوية باستخدام OneHotEncoder وOrdinalEncoder حيثما كان ذلك مناسبًا. يضمن ذلك أن جميع ميزات الإدخال كانت على نطاق قابل للمقارنة وبتنسيق مناسب لنماذج الانحدار.
- تم بعد ذلك تقسيم مجموعة البيانات إلى مجموعات تدريب واختبار باستخدام Train_test_split للسماح بالتدريب النموذجي والتقييم اللاحق. لزيادة تعزيز قوة عملية اختيار النموذج لدينا، استخدمنا StratifiedShuffleSplit لضمان الحفاظ على التقسيم الطبقي للبيانات، والحفاظ على توزيع المتغير المستهدف عبر الانقسامات.
- لاختيار النموذج، تم النظر في مجموعة متنوعة من خوارزميات الانحدار. وشملت هذه النماذج الخطية مثل LinearRegression وRidge وLasso وElasticNet، بالإضافة إلى نماذج أكثر تعقيدًا مثل RandomForestRegressor وGradientBoostingRegressor وDecisionTreeRegressor وSVR (Support Vector Regressor) وKNeighborsRegressor وXGBRegressor (eXtreme Gradient Boosting).
- لاستكشاف المعلمات الفائقة (hyperparameters) لهذه النماذج بشكل منهجي، تم استخدام GridSearchCV. وقد أتاح لنا ذلك إجراء بحث شامل حول قيم المعلمات المحددة لكل مقدر، وذلك باستخدام الأداء الذي تم التحقق من صحته لتحديد المعلمات الفائقة المثالية.
- كان اختيار الميزات جزءًا لا يتجزأ من خط أنابيبنا لتحسين قابلية تفسير النموذج والأداء. استخدمنا SelectFromModel لتحديد الميزات المهمة بناءً على أوزان أهميتها. وكانت هذه التقنية مفيدة بشكل خاص عند العمل مع النماذج التي يمكن أن توفر درجات أهمية الميزات، مثل الأساليب المستندة إلى الشجرة.

- ❖ تم تنظيم المسار باستخدام ColumnTransformer وPipeline من scikit-learn، مما يضمن تطبيق جميع خطوات المعالجة المسبقة بشكل متسق وتدريب النماذج بطريقة مبسطة. سهّل هذا النهج المعياري تجربة تقنيات المعالجة المسبقة المختلفة وخوارزميات الانحدار مع الحفاظ على سير عمل متماسك.
- ❖ تم إجراء تقييم النموذج باستخدام mean_absolute_error وr2_score لتقييم دقة النماذج وقوتها التفسيرية. قدمت هذه المقاييس فهمًا شاملاً لأداء النموذج، حيث يشير متوسط الخطأ المطلق إلى متوسط حجم الأخطاء ويعكس r2_score نسبة التباين التي يوضحها النموذج.
- ❖ وأخيرًا، تم حفظ النموذج الأفضل أداءً باستخدام joblib لاستخدامه في المستقبل، مما يسمح بالنشر الفعال وإجراء مزيد من التحليل.

في المرحلة الأولية من خط أنابيب المعالجة المسبقة للبيانات، قمنا بالعديد من الخطوات الرئيسية لإعداد مجموعة البيانات لمزيد من التحليل والنمذجة. تعتبر هذه الخطوات حاسمة لضمان جودة البيانات وسهولة استخدامها، خاصة في سياق التعلم الآلي.

1) حذف الأعمدة التي لا صلة لها

أولاً، قمنا بتحديد وإزالة عدة أعمدة من مجموعة البيانات التي اعتبرناها لا صلة لها بتحليلنا. تتضمن الأعمدة التي تم حذفها: salary_estimate, job_description, location, headquarters, founded, min_salary, max_salary

تم استبعاد هذه الأعمدة إما لأنها تحتوي على معلومات زائدة عن الحاجة، أو لم تكن مرتبطة بشكل مباشر بمهمة التنبؤ، أو اعتبرت غير ضرورية للتدريب النموذجي.

2) Encoding Categorical Features

بعد ذلك، ركزنا على تشفير الميزات الفئوية لتحويلها إلى تنسيق رقمي مناسب لخوارزميات التعلم الآلي. الميزات الفئوية التي تم تحديدها للتشفير هي: job_title, company_name, type_of_ownership, sector, job_city, job_state, hq_city, hq_state

3) الترميز الترتيبي

بالنسبة للميزات الفئوية المحددة التي لها ترتيب أو تصنيف متأصل، استخدمنا OrdinalEncoder لتشفير هذه القيم. تم تصميم التشفير للتعامل مع القيم غير المعروفة عن طريق تعيين قيمة معينة لها (-1 في هذه الحالة). وكانت أجهزة التشفير المحددة المستخدمة هي:

أداة تشفير الحجم: تم استخدام أداة التشفير هذه لتحويل معلومات حجم الشركة إلى قيم ترتيبية. تراوحت الفئات من "غير معروف" إلى "أكثر من 10000 موظف"، مما يسمح لنا بالتقاط التسلسل الهرمي للحجم.

أداة تشفير الإيرادات: يقوم هذا التشفير بتحويل فئات إيرادات الشركات إلى قيم ترتيبية. وتضمنت الفئات نطاقًا يتراوح بين "غير معروف / غير قابل للتطبيق" إلى "أكثر من 10 مليارات دولار (دولار أمريكي)".

أداة تشفير العمر: تم استخدام أداة التشفير هذه لتشفير عمر الشركات إلى قيم ترتيبية. وتراوحت الفئات من "غير معروف" إلى "قديم جدًا".

4) Binning Numerical Features

لتسهيل التعامل مع بيانات الرواتب وتحليلها بشكل أفضل، قمنا بإدراج عمود متوسط الراتب في فترات منفصلة. تم إنجاز ذلك باستخدام تابع pd.cut، التي قسمت متوسط الراتب إلى عدد محدد من الصناديق (num_bins = 10) وخصصت تسميات لها

5) أخذ العينات الطبقية وتقسيم اختبار التدريب

للتأكد من أن مجموعات التدريب والاختبار كانت ممثلة لمجموعة البيانات الإجمالية، لا سيما من حيث توزيع المتغير المستهدف (average_salary)، استخدمنا StratifiedShuffleSplit. تساعد هذه التقنية في الحفاظ على نسبة فئات المتغير المستهدف في كل من مجموعتي التدريب والاختبار.

6) إزالة عمود Binning

بعد التقسيم الطبقي، لم تعد هناك حاجة إلى عمود Average_salary_bin، الذي تم استخدامه فقط لغرض التقسيم الطبقي، وبالتالي تمت إزالته من مجموعتي التدريب والاختبار.

7) تقسيم الميزات والمتغير المستهدف

وكانت الخطوة التالية هي فصل الميزات (X) عن المتغير المستهدف (y). تم عزل المتغير المستهدف Average_salary، وتشكل الأعمدة المتبقية مجموعة الميزات.

8) إسقاط الأعمدة غير ذات الصلة من الميزات

تمت إزالة الأعمدة التي اعتبرناها بدون صلة بالنموذج (columns_to_drop) من كل من مجموعات التدريب والاختبار لتجنب أي ضجيج محتمل في البيانات.

9) تعريف محول العمود

للتعامل مع المعالجة المسبقة لأنواع مختلفة من الميزات بطريقة موحدة، تم تعريف ColumnTransformer. سمح لنا ذلك بتطبيق خطوات معالجة مسبقة مختلفة على مجموعات فرعية مختلفة من الميزات: القياس القياسي: يتم تطبيقه على الميزات الرقمية للتأكد من أن متوسطها 0 وانحراف معياري قدره 1. الترميز الساخن الواحد: يتم تطبيقه على الميزات الفئوية لتحويلها إلى تنسيق مناسب لخوارزميات التعلم الآلي. التشفير الترتيبي: يتم تطبيقه على ميزات فئوية محددة (الحجم، والإيرادات، وage_bin) باستخدام برامج تشفير ترتيبية محددة مسبقًا. الباقي: سيتم تمرير الأعمدة غير المحددة في المحول دون أي تغييرات.

10) تركيب وتحويل بيانات التدريب

تم بعد ذلك ملاءمة محول العمود لبيانات التدريب واستخدامه لتحويل مجموعة ميزات التدريب. وهذا يضمن أن جميع خطوات المعالجة المسبقة تم تطبيقها بشكل متسق عبر مجموعة البيانات.

11) One-Hot Encoding for Categorical Features

بالإضافة إلى ذلك، تم تطبيق One-Hot Encoding يدويًا على الميزات الفئوية في مجموعة التدريب لتحويلها إلى مصفوفة رقمية. أظهرت هذه الخطوة كيف يمكن تحويل الميزات الفئوية بشكل صريح.

قامت خطوات المعالجة المسبقة هذه بإعداد البيانات لمرحلة النمذجة اللاحقة، مما يضمن أن جميع الميزات كانت بتنسيق مناسب لخوارزميات التعلم الآلي.

12) اختيار الميزة عن طريق RandomForestRegressor

لتعزيز القدرة التنبؤية للنموذج وتقليل الأبعاد، استخدمنا خطوة اختيار الميزة باستخدام RandomForestRegressor. تعمل هذه الطريقة على الاستفادة من درجات أهمية الميزة المتأصلة التي توفرها خوارزمية Random Forest لتحديد الميزات الأكثر صلة.

تحديد نموذج اختيار الميزة: تم إنشاء RandomForestRegressor باستخدام 100 مُقدِّر وتم ضبطه لاستخدام جميع مراكز وحدة المعالجة المركزية المتاحة (n_jobs = -1). تم ضبط الحالة العشوائية للنموذج على 42 لضمان إمكانية التكاثر.

بناء خط الأنابيب: تم إنشاء خط الأنابيب لتبسيط خطوات المعالجة المسبقة واختيار الميزات. يتكون خط الأنابيب من: المعالجة المسبقة: استخدام ColumnTransformer (ct) المحدد مسبقًا لمعالجة البيانات مسبقًا. اختيار الميزة: استخدام SelectFromModel لإجراء اختيار الميزة بناءً على درجات الأهمية من RandomForestRegressor.

تركيب خط الأنابيب: تم بعد ذلك ملائمة خط الأنابيب لبيانات التدريب. يضمن ذلك تطبيق جميع خطوات المعالجة المسبقة واختيار الميزات الأكثر أهمية بناءً على الحد المتوسط.

استخراج الميزات المحددة: بعد تركيب المسار، تم تحديد الميزات المحددة باستخدام طريقة get_support الخاصة بـ SelectFromModel. قامت هذه الطريقة بإرجاع مؤشرات الميزات المحددة.

13) تحديد أسماء الميزات

لتفسير الميزات المحددة، تم استخراج أسماء الميزات المحددة. تضمنت هذه العملية دمج أسماء الميزات من خطوات المعالجة المسبقة المختلفة:

الميزات الرقمية: تم استخراج أسماء الميزات الرقمية من ColumnTransformer.

الميزات الفئوية: تم الحصول على أسماء الميزات الفئوية المشفرة الساخنة باستخدام طريقة get_feature_names_out.

الميزات الترتيبية: تمت إضافة أسماء الميزات الترتيبية مباشرةً حيث تم إدراجها بشكل صريح في ColumnTransformer.

الميزات المتبقية: تمت إضافة أي ميزات أخرى لم يتم تحويلها بشكل خاص.

14) إنشاء مجموعة الميزات النهائية

تم تجميع القائمة النهائية لأسماء الميزات المحددة من خلال توسيع قوائم أسماء الميزات الرقمية والفئوية والترتيبية والمتبقية

15) تحديد أفضل الميزات

واستنادا إلى عملية اختيار الميزات، تم تحديد مجموعة فرعية من أفضل الميزات ذات الصلة بالنموذج. تضمنت هذه القائمة مزيجًا من الميزات المتعلقة بالوظيفة والشركة والمهارات.

وتضمن هذه الخطوات اختيار الميزات الأكثر إفادة وإعدادها لمرحلة النمذجة اللاحقة، مما يعزز كفاءة وفعالية نموذج التعلم الآلي.

16) تقييم النموذج وضبط المعلمات الفائقة

لتحديد النموذج الأفضل أداءً لمهمة الانحدار لدينا، قمنا بتقييم العديد من النماذج باستخدام نهج خط الأنابيب الذي يتضمن المعالجة المسبقة واختيار الميزات والنموذج نفسه. لقد أجرينا أيضًا ضبط المعلمات الفائقة باستخدام GridSearchCV للعثور على المعلمات المثالية لكل نموذج.

تحديد شبكات المعلمات: لكل نموذج، حددنا نطاقًا من المعلمات الفائقة للبحث فيها. تضمنت هذه الشبكات مجموعات مختلفة من المعلمات لتحسين أداء النموذج.

تحديد النماذج: قمنا بدراسة مجموعة متنوعة من النماذج بما في ذلك النماذج الخطية والنماذج المبنية على الأشجار وطرق التجميع.

وظيفة التقييم: أنشأنا وظيفة لتقييم كل نموذج باستخدام GridSearchCV إذا تم توفير شبكة المعلمات. قامت هذه الوظيفة بتدريب النموذج وتوقع المتغير المستهدف وحساب مقاييس الأداء.

تقييم النموذج: قمنا بتقييم كل نموذج باستخدام الوظيفة المحددة وطبعنا النتائج، بما في ذلك أفضل المعلمات (إن أمكن)، والتدريب واختبار درجات R-squared، ومتوسط الخطأ المطلق (MAE)

أنتجت عملية التقييم النتائج التالية لكل نموذج، مع تفاصيل أدائها في مجموعات بيانات التدريب والاختبار:

1. **Linear Regression:**
   - **Train R-squared:** 0.8731
   - **Test R-squared:** 0.0729
   - **Train MAE:** 4.7468
   - **Test MAE:** 16.4348

كان أداؤه جيدًا في مجموعة التدريب ولكنه كان سيئًا في مجموعة الاختبار، مما يشير إلى overfitting المحتمل. الفرق الكبير بين التدريب واختبار MAE يدعم هذه الملاحظة.

2. **Ridge Regression:**
   - **Train R-squared:** 0.7828
   - **Test R-squared:** 0.3100
   - **Train MAE:** 8.3773
   - **Test MAE:** 14.8864

أظهر تعميمًا أفضل من الانحدار الخطي، مع وجود R-squared وMAE معقولين في مجموعة الاختبار. ومع ذلك، فإن فجوة الأداء بين مجموعات التدريب والاختبار تشير إلى درجة معينة من overfitting.

3. **Lasso Regression:**
   - **Train R-squared:** 0.1853
   - **Test R-squared:** 0.1985
   - **Train MAE:** 16.1906
   - **Test MAE:** 16.0036

كان أداؤه سيئًا، مما يشير إلى أنه قد لا يكون مناسبًا لمجموعة البيانات هذه. تشير قيم R-squared المنخفضة وMAE المرتفعة إلى أن النموذج يكافح من أجل التقاط الأنماط الأساسية.

4. **ElasticNet:**
   - **Best Params:** {'model__alpha': 0.1, 'model__l1_ratio': 0.9}
   - **Train R-squared:** 0.3319
   - **Test R-squared:** 0.3294
   - **Train MAE:** 14.7180
   - **Test MAE:** 14.7695

أظهر أداءً متواضعًا مع تعميم أفضل قليلاً مقارنةً بانحدار Lasso. تشير قيم R-squared وMAE إلى أن النموذج قد التقط بعض الأنماط ولكن ليس بشكل كافٍ

5. **Support Vector Regressor (SVR):**
    - **Best Params:** {'model__C': 1, 'model__epsilon': 0.5, 'model__kernel': 'linear'}
    - **Train R-squared:** 0.4024
    - **Test R-squared:** 0.3149
    - **Train MAE:** 12.6273
    - **Test MAE:** 13.9913

كان أداؤه جيدًا إلى حد معقول، مع قيم R-squared اللائقة وMAE. نجح النموذج في التعميم بشكل أفضل من بعض النماذج الخطية الأبسط.

6. **KNN Regression:**
    - **Train R-squared:** 0.3989
    - **Test R-squared:** 0.0843
    - **Train MAE:** 13.8572
    - **Test MAE:** 17.1532

أظهر تعميمًا ضعيفًا، مع انخفاض كبير في الأداء من التدريب إلى مجموعة الاختبار. يشير هذا إلى أن KNN قد لا يكون مناسبًا لمجموعة البيانات هذه.

7. **Decision Tree:**
    - **Best Params:** {'model__max_depth': 5, 'model__min_samples_split': 10}
    - **Train R-squared:** 0.3814
    - **Test R-squared:** 0.3190
    - **Train MAE:** 14.0801
    - **Test MAE:** 14.8239

كان أداؤه جيدًا إلى حد ما، مع أداء مماثل في كل من مجموعات التدريب والاختبار. وهذا يشير إلى توازن جيد بين التحيز والتباين.

8. **Random Forest:**
    - **Best Params:** {'model__max_depth': 10, 'model__min_samples_split': 2, 'model__n_estimators': 200}
    - **Train R-squared:** 0.6156
    - **Test R-squared:** 0.3883
    - **Train MAE:** 11.2893
    - **Test MAE:** 13.9015

أظهر أداءً جيدًا، مع قيم R-squared عالية نسبيًا وانخفاض MAE. أظهر هذا النموذج قدرة جيدة على التعميم.

9. **Gradient Boosting:**
    - **Best Params:** {'model__learning_rate': 0.1, 'model__max_depth': 5, 'model__n_estimators': 100}
    - **Train R-squared:** 0.5997
    - **Test R-squared:** 0.4019
    - **Train MAE:** 11.8327
    - **Test MAE:** 13.8298

كان أداؤه جيدًا، مع قيم R-squared جيدة وMAE منخفض. وأظهرت قدرات تعميم قوية.

17) نموذج XGBoost والاختيار النهائي

وبالنظر إلى أداء النماذج الأولية، ركزنا بعد ذلك على نموذج XGBoost المعروف بأدائه العالي في العديد من مهام الانحدار. لقد حددنا شبكة أنابيب ومعلمات منفصلة لـ XGBoost وأجرينا GridSearchCV لتحديد أفضل المعلمات الفائقة.

تحديد خط أنابيب XGBoost وشبكة المعلمات لـ XGBoost وبحث الشبكة وتقييم النموذج: تم إجراء GridSearchCV لـ XGBoost وطباعة أفضل المعلمات ونتائج التحقق المتبادل. التقييم النهائي لبيانات الاختبار: حساب متوسط الخطأ المطلق ومربع R على بيانات الاختبار.

B**est Parameters:**

- **Best Params:** {'model__learning_rate': 0.3, 'model__max_depth': 3, 'model__n_estimators': 50}
- **Best CV Score:** 14.6551

Test Evaluation:

- **Test MAE:** 13.9745
- **Test R-squared:** 0.3965

نلاحظ أنه في مجموعة الاختبار، حقق نموذج XGBoost متوسط MAE قدره 13.9745 و R-squared 0.3965، مما يشير إلى أداء قوي يمكن مقارنته بـ Random Forest وGradient Boosting.

6. المراجع:

[1] https://www.kaggle.com/code/nikolettaszab/data-anayst-job-description-word2vec

[2] https://www.kaggle.com/code/gawainlai/us-data-analyst-salary-exploratory-regression#More-on-California-Analyst-Salary-(Sector-&-Type-of-Ownership)

[3] https://www.kaggle.com/code/aienuefupi/data-analyst-jobs-eda-r#11.-Thanks~

[4] The investigation and prediction for salary trends in the data science industry {Wentao Jiang School of Science, Rensselaer Polytechnic Institute, 12180, USA}

[5] Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits – A Literature Review {Tee Zhen Quan Mafas Raheem}

[6] Salary Prediction Model using Principal Component Analysis and Deep Neural Network Algorithm

[7] https://www.kaggle.com/code/gawainlai/us-data-science-job-salary-regression-w-visuals

[8] https://www.kaggle.com/datasets/andrewmvd/data-analyst-jobs/data

[9] https://www.kaggle.com/code/labile10/ds-salary-cluster-and-regression

[10] https://www.kaggle.com/code/landfallmotto/data-analyst-jobs-eda-geolocation-map

[11] https://www.kaggle.com/code/kadirduran/beginner-friendly-eda-of-data-analyst-jobs

[12]https://www.kaggle.com/code/dileepbharati/data-analyst-jobs-eda#Highest-salary-jobs-by-location-and-rating

[13] https://www.kaggle.com/code/yusufglcan/eda-data-analyst-jobs

[14] SALARY PREDICTION USING REGRESSION TECHNIQUES Sayan Das(JIS College of Engineering, Kalyani, Nadia), Rupashri Barik*(JIS College of Engineering, Kalyani, Nadia), Ayush Mukherjee(JIS College of Engineering, Kalyani, Nadia).

[15] A Study on Recruitment of Data Analyst Based on Text Mining and Visualization Technology

{Fang Fang1, *, Yin Zhou1, 2}