

Metodologia Analitică a Pulsului Tehnologic: Evaluarea Statistică a Salariilor, Distribuției Geografice și Impactului Muncii la Distanță în Sectorul IT

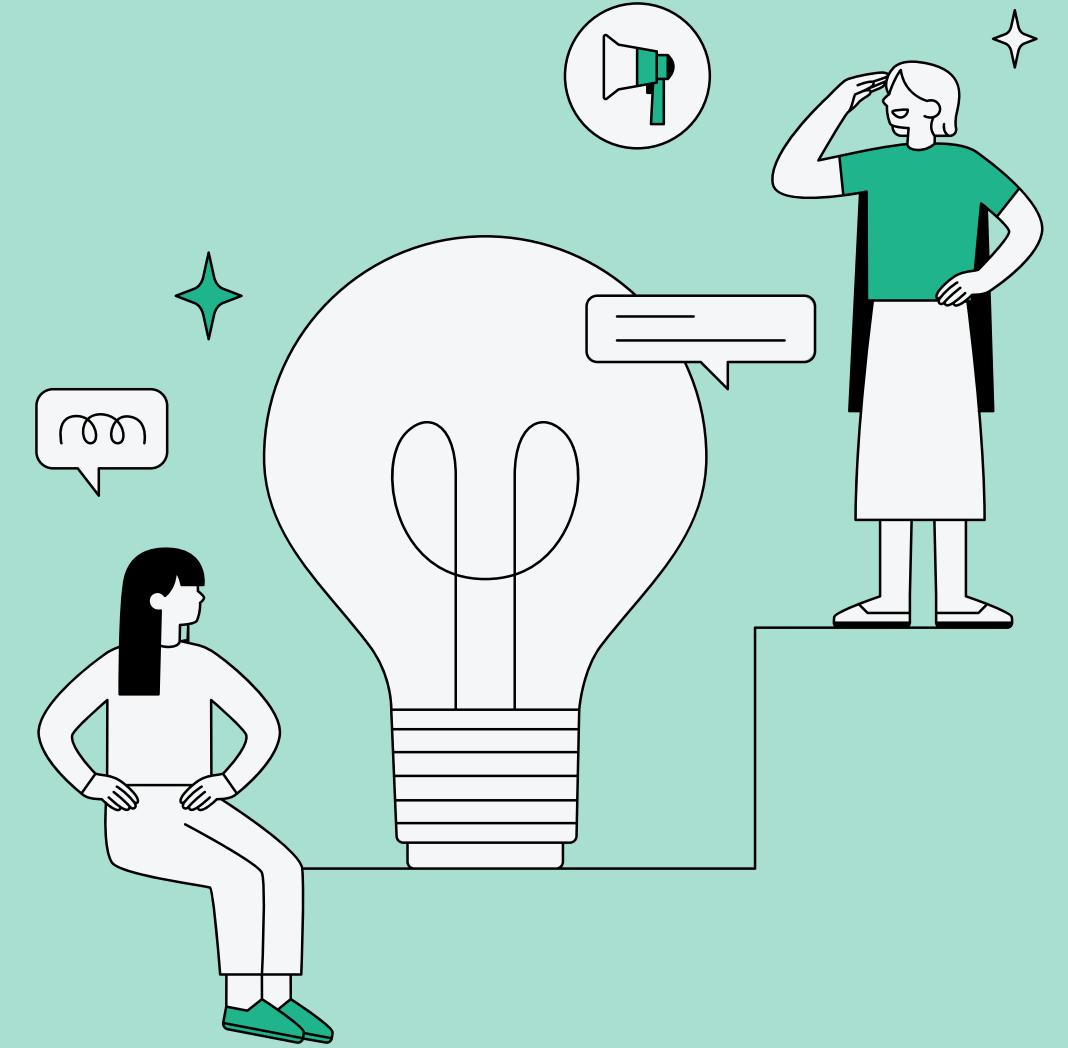
<https://github.com/CorneliuMagurean/Proiect-Analiza-Salariilor-Analiza-Datelor>

Magurean Corneliu, gr. MI-211



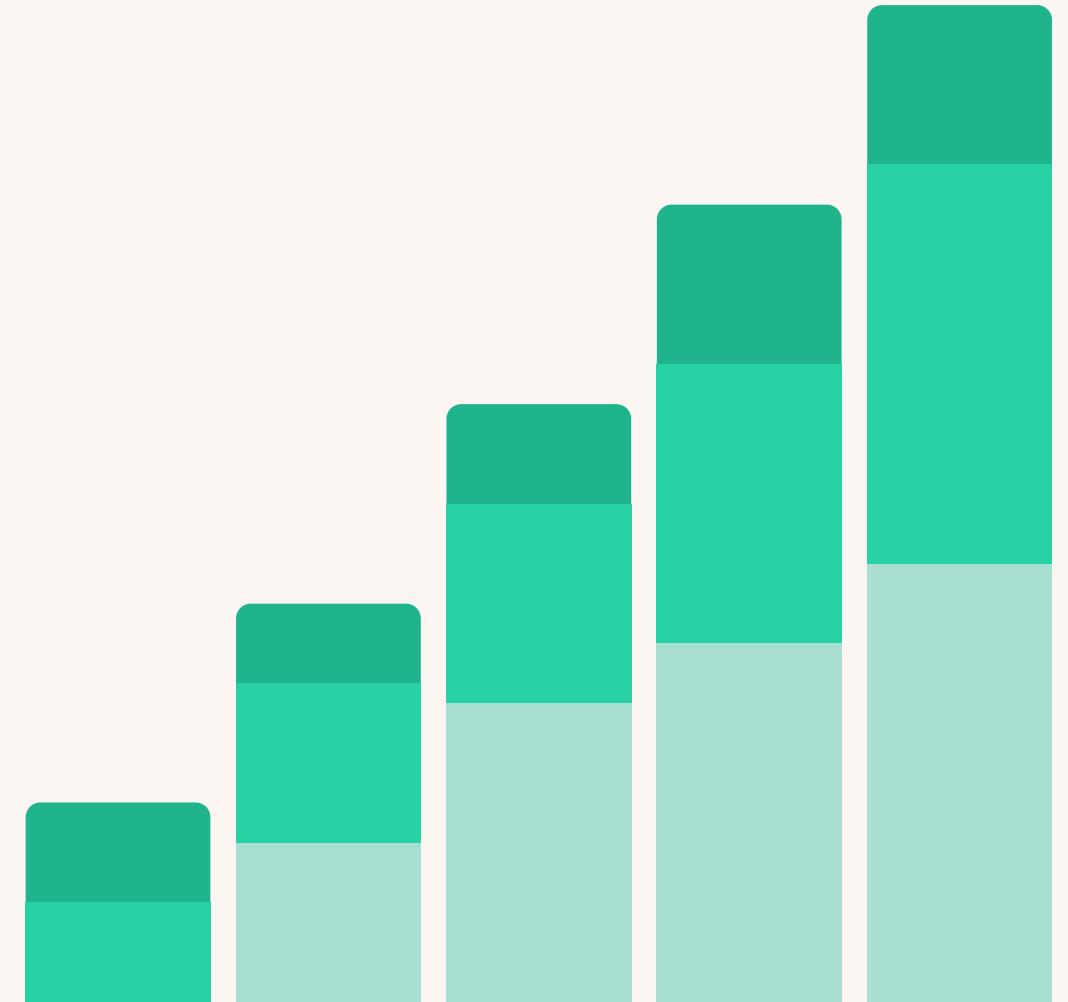
Introducere

Tehnologia și digitalizarea au reconfigurat peisajul profesional, cu sectorul IT în fruntea acestei transformări. Salariile și modul în care munca la distanță influențează acestea au devenit subiecte centrale în discuții. "Metodologia Analitică a Pulsului Tehnologic: Evaluarea Statistică a Salariilor, Distribuției Geografice și Impactului Muncii la Distanță în Sectorul IT" se concentrează pe analiza riguroasă a acestor tendințe. Vom examina datele salariale din diverse regiuni, evaluând impactul variabilelor precum locația și experiența, și vom investiga rolul muncii remote în modelarea compensației în domeniu.



Problematica

În dinamica rapidă a sectorului IT, salariile și structura lor constituie un barometru al valorii, cererii și evoluției profesionale. Cu toate acestea, nu există doar o singură măsură universală; discrepanțele salariale pot fi influențate de variații geografice, culturale sau economice, creând un mozaic complex de compensații. Adăugând la această complexitate, tranziția accelerată către munca la distanță aduce noi provocări și oportunități. Cum modeleză munca remote structura salarială? Există avantaje sau dezavantaje salariale pentru cei care optează pentru flexibilitate? Această proiect își propune să discearnă aceste intersecții și să ofere o perspectivă clarificatoare asupra salariilor în era digitală.



Challenge-uri

Industria IT, în ciuda avansului său rapid și aportului semnificativ la economie, prezintă o serie de provocări care necesită atenție și analiză. Printre acestea:

1. Discrepanțele salariale: De ce anumiți profesioniști sunt plătiți mai mult decât alții pentru același rol, și cum influențează factori precum locația, experiența sau educația aceste diferențe?
2. Munca la distanță: Odată cu creșterea muncii remote, cum sunt afectate salariile și oportunitățile profesionale? Există vreo corelație între flexibilitatea muncii și compensație?
3. Evoluția sectorului: Cu tehnologiile emergente și schimbările în cererea de competențe, cum se reflectă acestea în structura salarială?



Motivația Analizei

Această analiză este crucială din mai multe motive:

- Informare: Oferă o perspectivă clară asupra structurii salariale în sectorul IT, ajutând angajatorii să ia decizii informate în strategiile lor de compensație.
- Planificarea carierei: Profesioniștii IT pot înțelege mai bine unde se situează pe piață și ce oportunități există pentru creșterea lor profesională și salarială.
- Tendințe și evoluții: Înțelegerea modului în care industria se schimbă și impactul acestor schimbări asupra compensației poate oferi o viziune asupra direcției în care se îndreaptă sectorul IT în viitor.

Analiza Setului de Date și a Variabilelor

1. Descrierea Setului de Date

Setul de date pe care îl avem la dispoziție pare să conțină informații despre salariile profesioniștilor IT din diferite regiuni, cu diverse niveluri de experiență și în diferite poziții.

2. Variabilele și Semnificația Lor

- **work_year:** Aceasta ne indică anul în care a fost înregistrat salariul, oferindu-ne o bază temporală pentru analiză.
- **experience_level:** Nivelul de experiență este adesea corelat cu salariul, așteptându-ne ca persoanele cu mai multă experiență să aibă salarii mai mari.
- **employment_type:** Tipul de angajare (full-time, contract) poate influența mărimea și structura compensației.
- **job_title:** Poziția sau rolul profesional poate avea un impact semnificativ asupra salariului.
- **salary și salary_currency:** Acestea ne oferă informații despre compensație în moneda locală.
- **salary_in_usd:** Convertirea salariului în USD ne permite să facem comparații standardizate între țări.
- **employee_residence și company_location:** Locația poate influența salariul datorită costului vieții și cererii de competențe în diferite regiuni.
- **remote_ratio:** Procentul de muncă la distanță poate influența salariul, în special în contextul actualelor tendințe de lucru remote.
- **company_size:** Mărimea companiei poate reflecta bugetul disponibil pentru salarii și beneficii.



Analiza Exploratorie a Datelor (EDA)

Analiza Valorilor Lipsă

Verificăm fiecare coloană din setul de date pentru a identifica valori lipsă și sumăm numărul lor.

work_year	experience_level	employment_type	job_title	salary
0	0	0	0	0
salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location
0	0	0	0	0
company_size				
0				

În analiza noastră preliminară, am examinat prezența valorilor lipsă în setul de date al salariilor din sectorul IT. Valorile lipsă pot influența integritatea și calitatea analizei noastre, de aceea este esențial să le identificăm în fazele incipiente.

Din fericire, rezultatele noastre indică faptul că setul de date este complet, fără nicio valoare lipsă în niciuna dintre coloane. Acest lucru sugerează că avem un set de date bine curățat și pregătit pentru analiză.



Statistică Descriptive

Statisticile descriptive oferă o privire generală asupra distribuției și tendințelor centrale ale datelor. Acestea sunt esențiale pentru a obține o înțelegere preliminară a setului de date.

```
> # Afisarea statisticilor
> print(summary_stats_continuous)
  mean_salary median_salary sd_salary min_salary max_salary
1    190695.6        138000   671676.5       6000  30400000
>
> # Frecvența pentru variabilele categorice (experience_level)
> experience_level_freq <- table(salaries$experience_level)
>
> # Afisarea frecvențelor
> print(experience_level_freq)

EN   EX   MI   SE
320  114  805 2516
```

Analizând setul de date al salariilor din sectorul IT, am obținut următoarele observații:

1. Distribuția Salariilor:

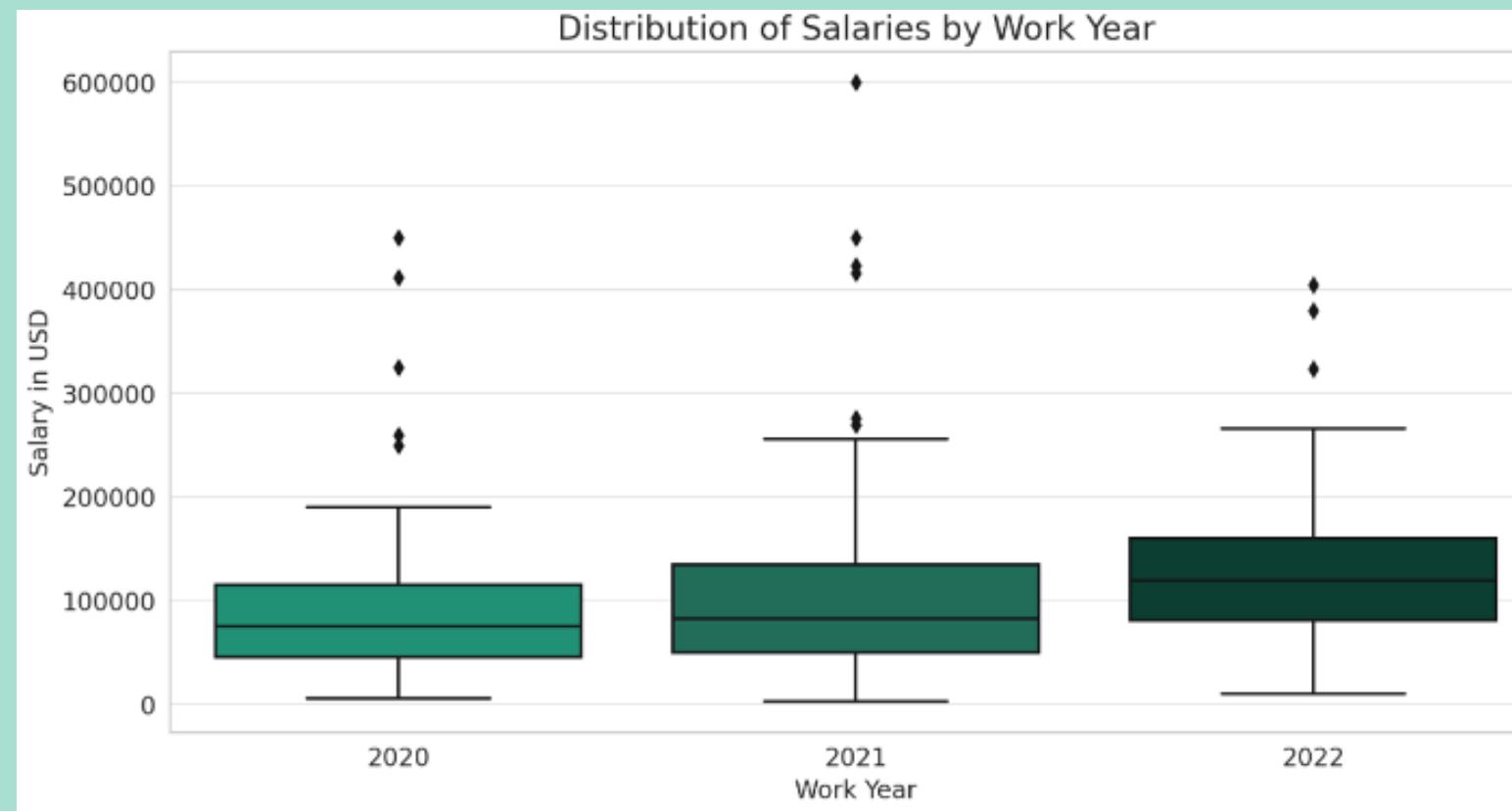
- Media salarială: Media salariilor este de aproximativ 190,696 unități, sugerând că în medie, un angajat din sectorul IT câștigă această sumă.
- Mediana salarială: Mediana, sau salariul de mijloc, este de 138,000 unități, indicând că jumătate dintre angajați câștigă sub această sumă, iar cealaltă jumătate peste.
- Deviația Standard: Cu o valoare de 671,676.5, deviația standard este destul de mare, indicând o variație semnificativă în salariile angajaților.
- Salariul Minim și Maxim: Salariile variază considerabil, de la un minim de 6,000 unități la un maxim uimitor de 30,400,000 unități.

2. Nivelul de Experiență:

- Entry Level (EN): Există 320 de angajați la nivel de început în setul de date.
- Experienced (EX): 114 angajați au fost clasificați ca având un nivel de experiență.
- Middle (MI): Cu 805 angajați, nivelul mediu este bine reprezentat.
- Senior (SE): Majoritatea angajaților, 2,516 la număr, sunt la un nivel superior, ceea ce indică faptul că setul de date este dominat de profesioniști cu o vastă experiență în domeniu.

Distributia salariilor pe ani:

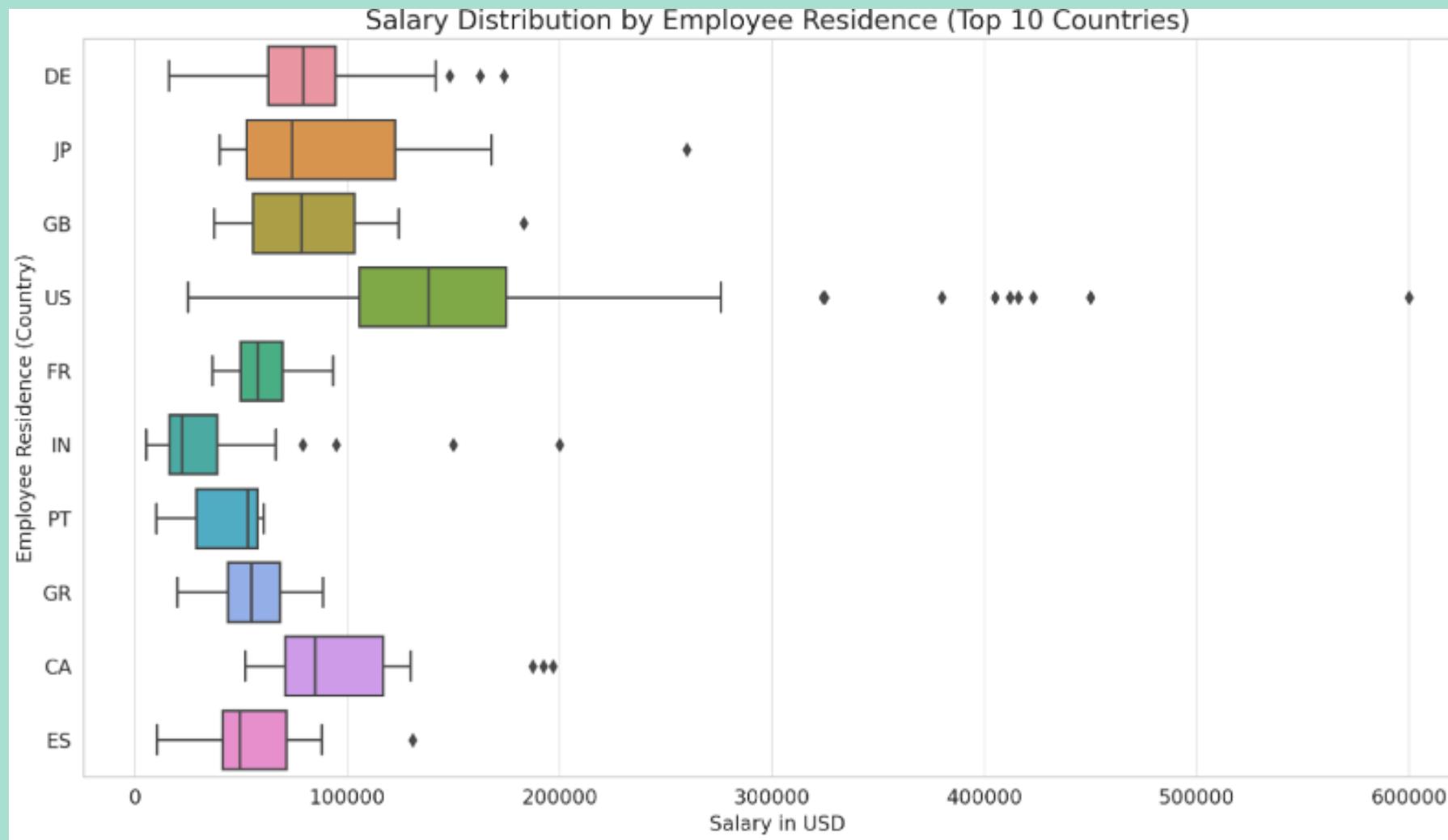
Vom vizualiza distribuția salariilor în dolari americani (USD) pentru a înțelege variația și tendința centrală.



Există o variație a salariilor în fiecare an, cu o tendință de creștere a medianei și a intervalelor intercartilice pe măsură ce avansăm de la 2020 la 2022. Aceasta poate indica o creștere generală a salariilor în sectorul IT în această perioadă.

- Creștere Anuală: S-a observat o tendință de creștere a salariilor medii și mediane de la anul 2020 până în 2022. Aceasta sugerează că industria IT a experimentat o creștere salarială în această perioadă, posibil datorită creșterii cererii de competențe tehnologice și a evoluției pieței muncii în acest sector.

Distributia salariilor in functie de tara de reședinta



Există diferențe semnificative în salariile angajaților în funcție de țara lor de reședință. Angajații din SUA au, în general, cele mai înalte salarii, urmați de cei din CA (Canada) și GB (Marea Britanie). Salariile din alte țări, cum ar fi IN (India) și BR (Brazilia), sunt mai scăzute. Aceasta reflectă probabil diferențele economice și de piață între aceste regiuni.

- Discrepanțe Regionale: Observăm diferențe semnificative în salarii în funcție de reședința angajaților. SUA se distinge prin salarii considerabil mai ridicate comparativ cu alte țări. Acest lucru ar putea reflecta nivelul de dezvoltare economică, costul vieții și concentrarea industriei IT în SUA.

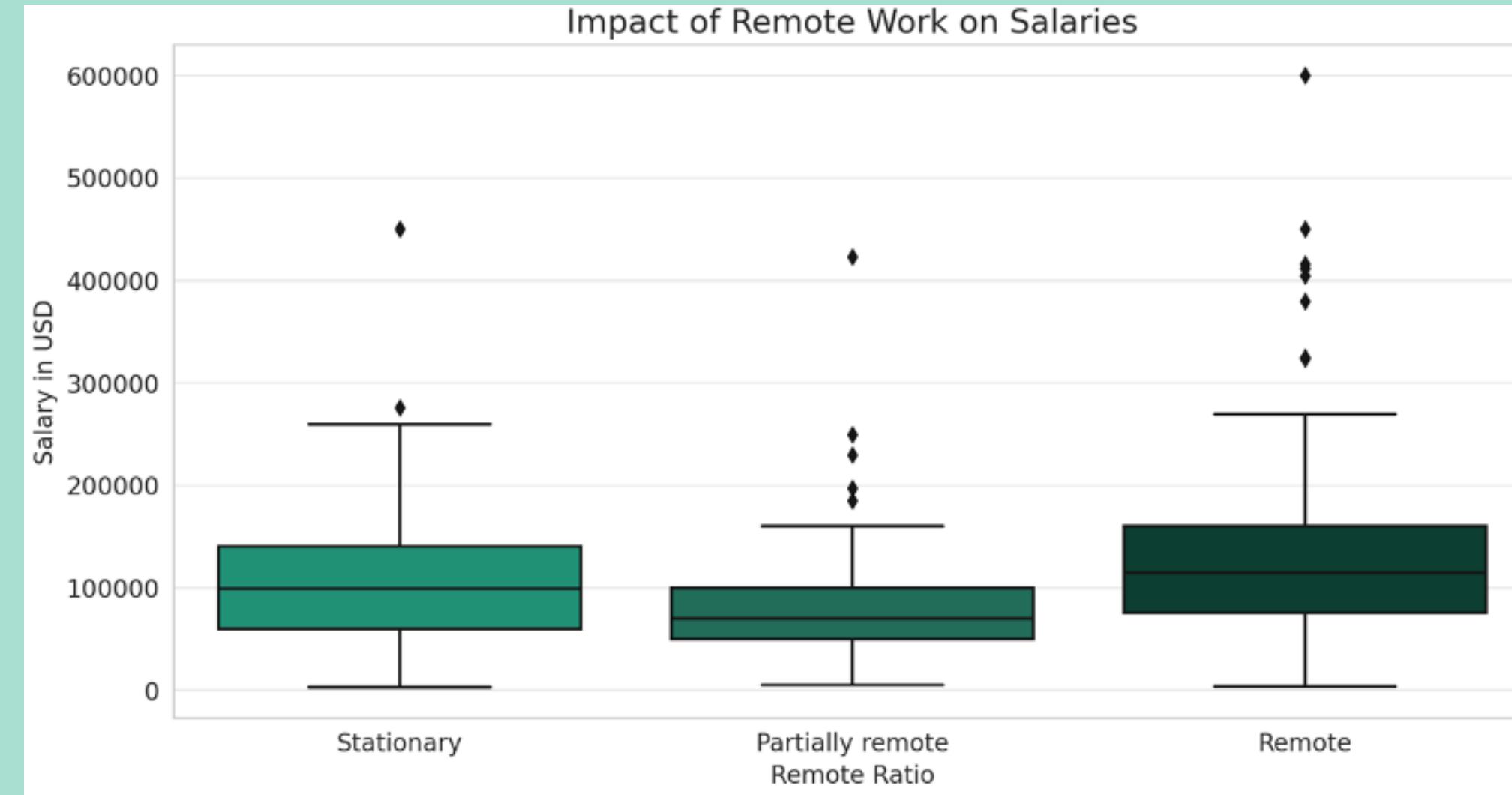
Distributia salariilor pe nivel de experienta



Salariile tend să crească odată cu nivelul de experiență. "Senior/Expert" are cel mai înalt intercartilic, indicând salarii mai mari pentru acest grup. "Entry-level" are cele mai scăzute salarii, ceea ce este de așteptat.

- Diferențe în Funcție de Experiență: Salariile cresc semnificativ odată cu nivelul de experiență. Angajații la nivelul "Senior/Expert" au cele mai mari salarii, indicând o valoare mare a experienței și expertizei în domeniu. În contrast, salariile la nivel "Entry-level" sunt considerabil mai scăzute, ceea ce este caracteristic pentru profesioniștii aflați la început de carieră.

Impacul lucrului la distanță asupra salariului

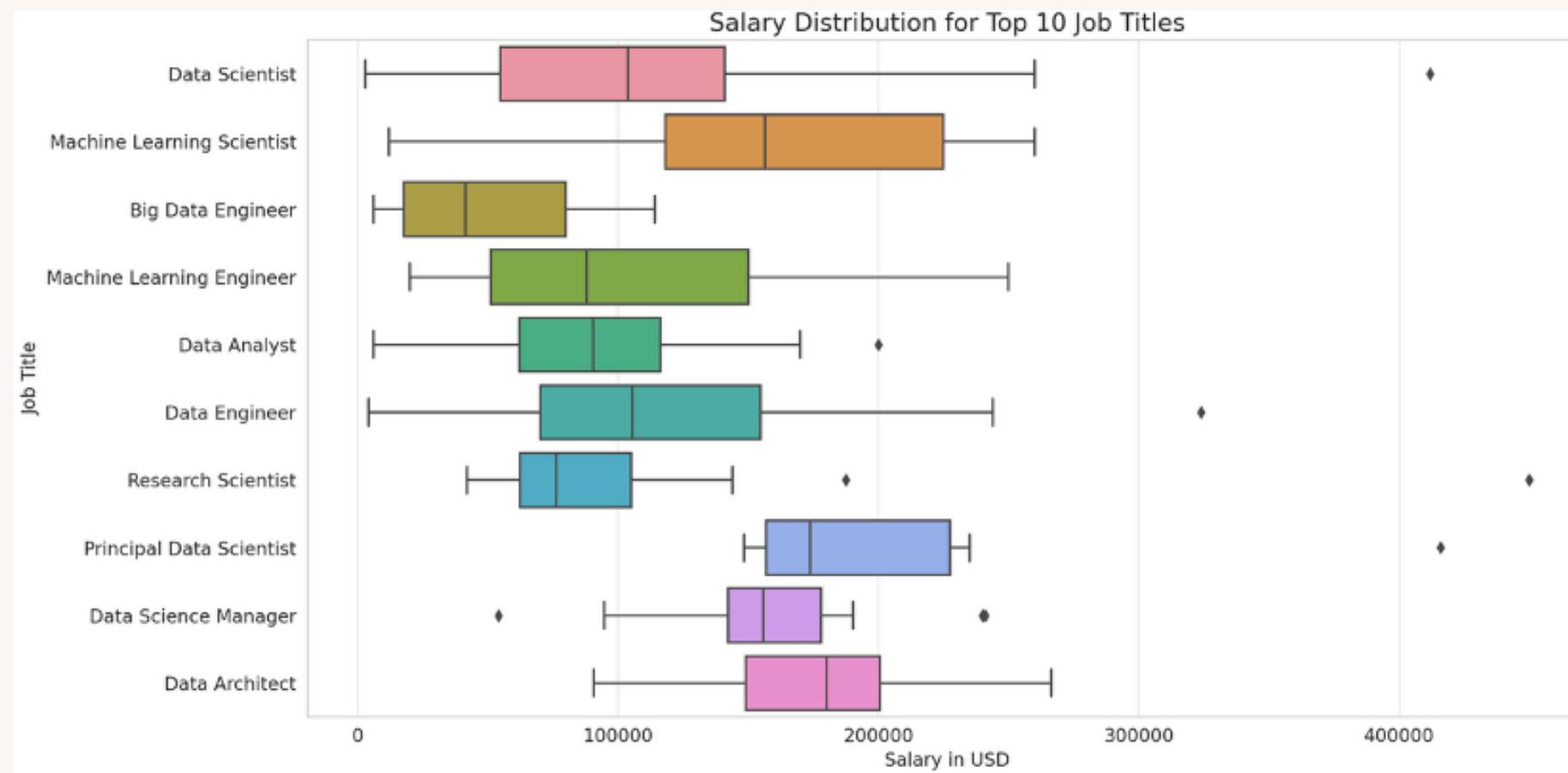


Se observă că există diferențe în salariile oferite în funcție de raportul de muncă la distanță. Angajații care lucrează complet la distanță ("Remote") par să aibă salarii mediane mai mari comparativ cu cei care lucrează parțial sau deloc la distanță.

Angajații care lucrează parțial la distanță ("Partially remote") au o distribuție a salariilor similară cu cei care nu lucrează la distanță ("Stationary").

Acstea descoperiri sugerează că muncirea complet la distanță poate fi asociată cu salarii mai mari, posibil datorită accesului la o piață de muncă mai largă și a flexibilității în alegerea locurilor de muncă.

Distributia salariilor în funcție de job



Există o variație semnificativă a salariilor în funcție de titlul postului. Anumite titluri, cum ar fi "Machine Learning Scientist" și "Data Engineer", par să aibă salarii mai ridicate, în timp ce altele, cum ar fi "Data Analyst", au salarii mai scăzute. Aceasta reflectă diferențele de cerere și specializare în diferite roluri.

- Variabilitate Mare între Roluri: Diferențele salariale între diversele titluri de post sunt remarcabile. Anumite roluri, cum ar fi "Machine Learning Scientist" și "Data Engineer", sunt asociate cu salarii semnificativ mai ridicate, reflectând cererea mare pentru aceste competențe specializate în piața muncii.

Testarea ipotezelor specifice bazate pe observațiile și descoperirile făcute în EDA

1. Impactul locației asupra salariilor: Se va examina dacă există diferențe semnificative în salariile angajaților IT bazate pe locația geografică, atât din perspectiva țării de reședință a angajaților, cât și a locației companiilor.
2. Relația dintre nivelul de experiență și salariu: Se va analiza dacă un nivel mai înalt de experiență este asociat cu salarii mai mari, sugerând o valorizare a experienței acumulate în acest sector.
3. Variabilitatea salariilor în funcție de titlul postului: Se va investiga cum influențează diferitele roluri și responsabilități din IT structura salarială.
4. Impactul muncii la distanță asupra salariilor: Se va explora cum diferitele moduri de muncă la distanță (complet, parțial sau fără muncă la distanță) influențează remunerația angajaților.

Impactul locației asupra salariilor

- Rezultate: Statistica F de 6.34 și o p-valoare de 1.06e-32 indică diferențe semnificative în salariile angajaților IT în funcție de locația geografică.
- Interpretare: Aceste rezultate confirmă ipoteza că locația angajaților și a companiilor are un rol crucial în determinarea salariilor în sectorul IT. Diferențele semnificative între locații sugerează că factori precum economia locală, costul vieții și cererea de competențe specifice influențează remunerația.

Relația dintre nivelul de experiență și salariu

- Rezultate: O statistică F de 64.68 și o p-valoare de 2.88e-36 arată că există diferențe semnificative în salarii între diferitele niveluri de experiență.
- Interpretare: Aceste descoperiri susțin ipoteza că nivelul de experiență este un factor determinant în structura salarială din IT. Salariile mai mari la nivelurile superioare de experiență reflectă valoarea experienței acumulate și a competențelor avansate.

Variabilitatea salariilor în funcție de titlul postului

- Rezultate: Statistica F de 3.81 și o p-valoare de 6.89e-15 indică variații semnificative în salarii în funcție de titlul postului.
- Interpretare: Acest rezultat validează ipoteza că titlul postului este un indicator important al structurii salariale. Variațiile semnificative între diferite roluri sugerează că anumite competențe sau responsabilități sunt mai valorizate pe piața muncii.

Impactul muncii la distanță asupra salariilor

- Rezultate: O statistică F de 14.73 și o p-valoare de 5.68e-07 demonstrează impactul semnificativ al muncii la distanță asupra salariilor.
- Interpretare: Acest rezultat confirmă ipoteza că munca la distanță influențează structura salarială. Faptul că angajații care lucrează complet la distanță tind să aibă salarii medii mai mari arată cum flexibilitatea și accesul la o piață de muncă mai largă pot influența remunerația.

Modele

Un aspect esențial în analiza datelor și înțelegerea fenomenelor complexe, cum ar fi structura salarială în sectorul IT, este utilizarea modelelor statistice și analitice. În acest capitol, se va explora diferite modele de regresie și analiză statistică aplicate pe setul de date, pentru a evalua și cuantifica impactul diferenților factori asupra salariilor. Scopul acestor modele este de a oferi o înțelegere mai profundă a relațiilor dintre variabile și de a genera predicții precise și relevante.

Se vor utiliza mai multe tipuri de modele, fiecare cu specificările și avantajele sale, pentru a aborda ipotezele formulate în studiu. Aceasta include:

1. Regresia Liniară: Un model de bază, care va servi la evaluarea relațiilor liniare dintre variabile.
2. Regresia Ridge: O extensie a regresiei liniare, utilă în gestionarea multicolinearității și în îmbunătățirea preciziei modelului.
3. Arborele de Decizie pentru Regresie: O abordare non-liniară, care va fi explorată pentru a capta relații mai complexe și interacțiuni între variabile.

Regresie liniară

Rezumatul modelului

- R-squared: 0.485
 - Aproximativ 48.5% din variația salariilor poate fi explicată de variabilele independente incluse în model. Acesta este un nivel moderat de explicabilitate.
- Adjusted R-squared: 0.422
 - Aceasta este un indicator mai precis pentru modelele cu mai multe variabile independente și sugerează că modelul se adaptează rezonabil de bine la date.
 -

Observații importante

- Multicolinearitate: Există o posibilă problemă de multicolinearitate, aşa cum este indicat de notele modelului. Acest lucru poate afecta fiabilitatea coeficienților individuali.
- Variabile semnificative: Anumite variabile, cum ar fi nivelul de experiență, par să aibă un impact semnificativ asupra salariilor, aşa cum este indicat de coeficienți.
- Variabilitatea variabilelor categorice: Datorită numărului mare de țări incluse, este posibil ca unele să nu aibă un număr suficient de mare de observații pentru a oferi o estimare precisă.

Model de regresie Ridge

Evaluarea modelului

- R-squared: 0.499
 - Modelul poate explica aproximativ 49.9% din variația salariilor. Acest lucru indică o îmbunătățire față de modelul anterior și o ajustare rezonabilă a datelor.
- RMSE (Root Mean Square Error): 43800.97 USD
 - Aceasta este eroarea medie pătratică a predicțiilor modelului. Este o măsură a discrepanței dintre valorile reale ale salariilor și predicțiile modelului. O valoare mai mică a RMSE indică o precizie mai mare.

Concluzii și recomandări

- Modelul Ridge oferă o analiză mai robustă a datelor, mai ales în prezența multicolinearității. Aceasta poate fi o abordare mai adecvată pentru seturile de date complexe, cum ar fi cel analizat.
- Interpretarea coeficienților și a importanței relative a variabilelor ar fi un pas următor valoros pentru a înțelege mai bine care factori influențează cel mai mult salariile în sectorul IT.

Este important să se țină cont de limitările modelului, inclusiv de posibilele variabile omise sau de biasul inherent în date, înainte de a trage concluzii finale.

Modelului de regresie bazat pe arbore de decizie

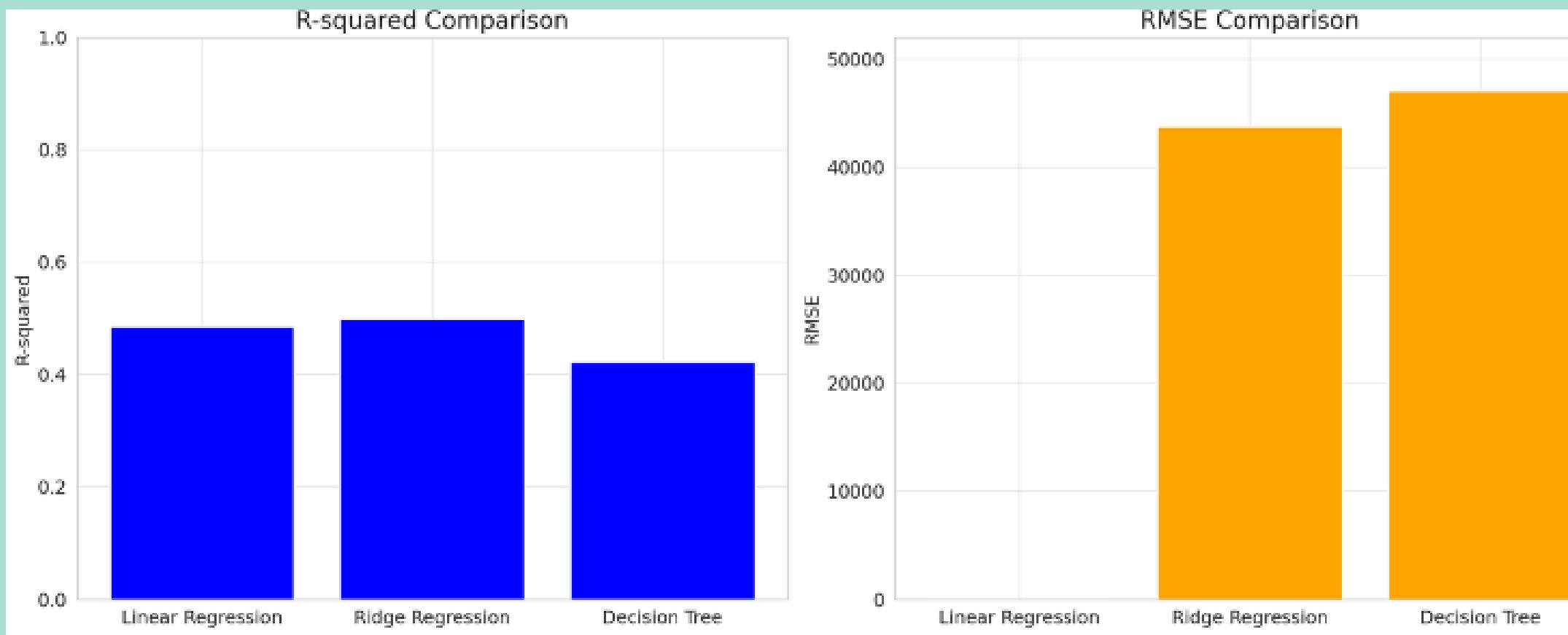
Evaluarea modelului

- R-squared: 0.423
- Modelul poate explica aproximativ 42.3% din variația salariilor. Acesta este un nivel moderat de explicabilitate, inferior modelului Ridge.
- RMSE (Root Mean Square Error): 47035.39 USD
- Aceasta este eroarea medie pătratică a predicțiilor modelului. O valoare mai mare a RMSE față de modelul Ridge indică o precizie mai scăzută.

Concluzii și recomandări

- Modelul de arbore de decizie oferă o alternativă la modelele liniare, dar în acest caz, nu pare să fie cea mai bună alegere pentru setul nostru de date.
- S-ar putea lua în considerare utilizarea unor tehnici de regularizare și ajustare a hiperparametrilor (de exemplu, prin limitarea adâncimii arborelui) pentru a îmbunătăți performanța modelului de arbore de decizie.
- Este important să se țină cont de faptul că modelele mai complexe nu sunt întotdeauna cele mai bune; uneori, un model mai simplu, dar bine ajustat, poate oferi rezultate mai robuste și mai ușor de interpretat.

Comparare modele



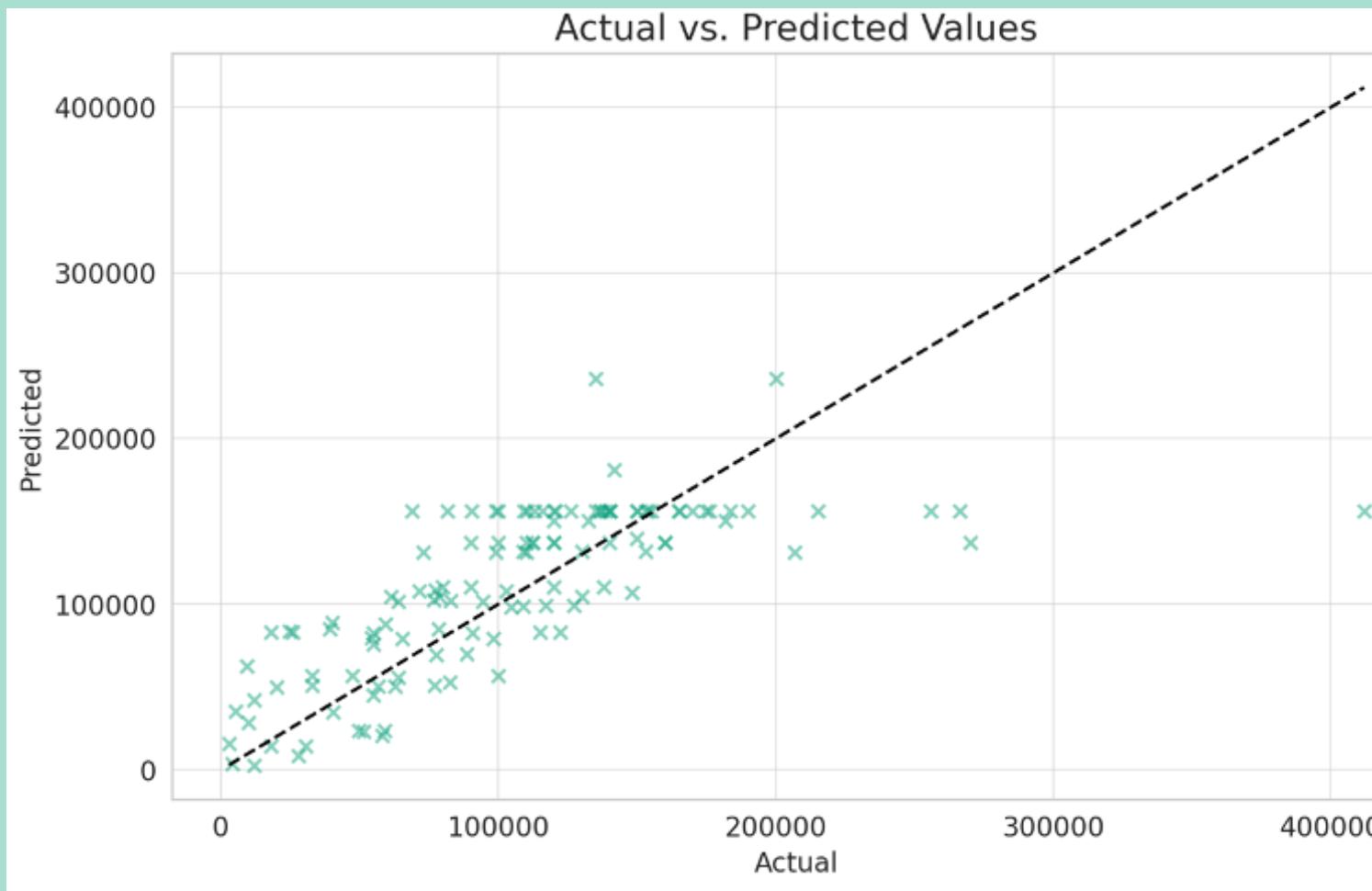
Concluzii:

- Modelul Optimal: Bazându-ne pe aceste două metri, regresia Ridge pare să fie modelul cel mai performant pentru setul de date analizat. Acesta nu doar că explică o proporție mai mare din variația salariilor, dar are și o precizie mai mare în predicții.
- Recomandări: În contextul acestui set de date și pentru scopul analizei - evaluarea influenței locației, nivelului de experiență și muncii la distanță asupra salariilor - regresia Ridge ar fi cea mai potrivită alegere.

Aceste rezultate sugerează că, în cazul în care căutați un model care să ofere atât o bună explicabilitate, cât și o predicție precisă, regresia Ridge este cea mai adecvată opțiune dintre cele evaluate.

Modelul optimal – Ridge

Plotul de dispersie al valorilor prezise vs. valorilor reale



Interpretare:

- **Concentrația Punctelor:** O concentrare a punctelor în jurul liniei punctate arată că modelul face predicții precise. Dacă punctele sunt împrăștiate larg, arată o variație mare între valorile prezise și cele reale.
- **Outliers:** Punctele care se abat semnificativ de la linia punctată pot indica cazuri unde modelul nu a funcționat bine.
- **Tendința Generală:** Dacă majoritatea punctelor se aliniază relativ bine pe linie, sugerează că modelul are o performanță bună.

Concluzii

Rezultatele studiului au validat majoritatea ipotezelor inițiale. S-a confirmat că locația geografică, nivelul de experiență, titlul postului și gradul de muncă la distanță sunt factori cheie care influențează salariile în sectorul IT. Această constatare subliniază importanța unei perspective globale și adaptate la realitățile locale în gestionarea carierei și politicilor organizaționale.

Implicații pentru Industria IT

- Pentru profesioniști: Înțelegerea acestor factori este crucială pentru navigarea eficientă pe piața muncii, planificarea carierei și negocierile salariale.
- Pentru organizații: Rezultatele studiului oferă informații valoroase pentru strategii de recrutare, retenție și dezvoltare a talentelor, precum și pentru formarea unor structuri salariale echitabile și competitive.

