Introducere

În lumea modernă a recrutării și gestionării resurselor umane, este crucial să evaluăm angajabilitatea și gradul de salarizare al indivizilor pentru a lua decizii informate și eficiente.’ În acest context, seturile de date bogate și variate devin instrumente esențiale pentru înțelegerea dinamicilor pieței muncii și pentru identificarea tendințelor relevante.

Setul nostru de date cuprinde informații detaliate despre angajați, inclusiv nivelul lor de educație, anul în care s-au alăturat organizației, orașul de rezidență, salarizarea, vârsta, genul, experiența în domeniul actual, precum și informații despre dacă au fost sau nu vreodată înlocuiți și dacă au părăsit sau nu organizația.

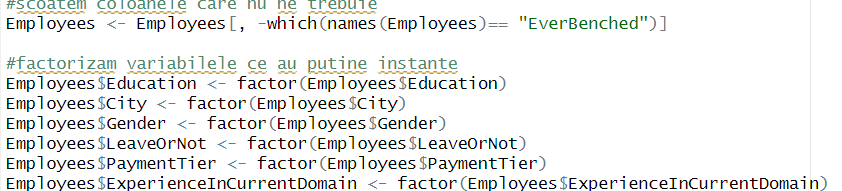
Atributele din setul de date sunt următoarele: Education (Educație): Indică nivelul de educație al angajatului și poate fi "Bachelors" (Licență), "Masters" (Masterat) sau "PhD" (Doctorat). Aceasta este o variabilă nominală. Joining Year (Anul Alăturării): Reprezintă anul în care angajatul s-a alăturat organizației. Este o variabilă numerică. City (Oraș): Indică orașul de reședință al angajatului. Este o variabilă nominală. PaymentTier (Nivelul de Salarizare): Arată nivelul de salarizare al angajatului, exprimat printr-o valoare numerică. Este variabila dependentă pe care ne propunem să o analizăm în raport cu celelalte atribute. Age (Vârsta): Indică vârsta angajatului exprimată în ani. Este o variabilă numerică. Gender (Genul): Reprezintă genul angajatului și poate fi "Male" (Masculin) sau "Female" (Feminin). Este o variabilă nominală. EverBenched (Vreodată Înlocuit): Indică dacă angajatul a fost vreodată înlocuit ("Yes" sau "No"). Este o variabilă nominală. Experience (Experiență): Arată nivelul de experiență al angajatului în domeniul actual exprimat în ani. Este o variabilă numerică. Leave or Not (Părăsire sau Nu): Indică dacă angajatul a părăsit organizația sau nu ("1" pentru "Da" și "0" pentru "Nu"). Este o variabilă nominală.

Întrebările noastre de cercetare sunt: **Există o corelație între nivelul de salarizare al unui angajat și vârsta acestuia? Există o relație între nivelul de salarizare al unui angajat și experiența sa în domeniul actual?**

Aceste întrebări sunt deosebit de importante pentru managementul resurselor umane și pentru departamentele de HR, deoarece răspunsurile pot oferi o perspectivă mai clară asupra politicii de salarizare, pot ajuta la identificarea posibilelor discrepanțe sau neajunsuri în procesele de recrutare și pot contribui la dezvoltarea strategiilor mai eficiente de retenție a personalului. Prin analiza și înțelegerea relațiilor dintre aceste variabile, organizațiile pot lua decizii mai informate și pot optimiza procesele lor de recrutare și gestionare a resurselor umane.

Factorizarea datelor

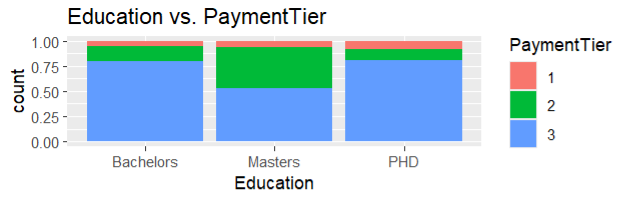
În continuare, am realizat o prelucrare a setului de date "Employees" ( ) pentru a-l pregăti pentru analiza ulterioară. Am eliminat coloana "EverBenched" din setul de date și am transformat mai multe variabile în factori pentru a le trata ca categorii distincte în analizele ulterioare. Această transformare ne permite să lucrăm mai eficient cu datele nominale și să obținem rezultate mai precise în analizele statistice și vizualizările datelor.



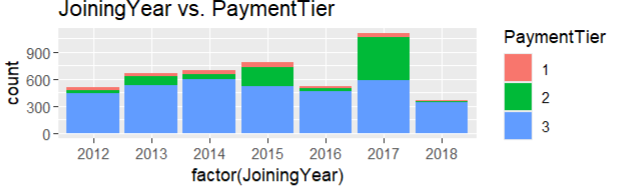
Vizualizări ale Distribuției Nivelului de Salarizare în Funcție de Caracteristicile Angajaților

Aceste vizualizari sunt utilizate pentru a crea o serie de grafice folosind pachetul **ggplot2**. Fiecare grafic compară diferite caracteristici ale angajaților din setul de date "Employees" în funcție de nivelul lor de salarizare ("PaymentTier").

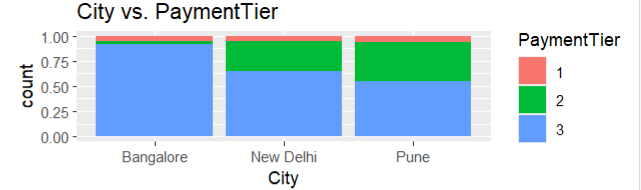
**1.** **plot1**: Afișează distribuția nivelului de salarizare în funcție de nivelul de educație al angajaților.



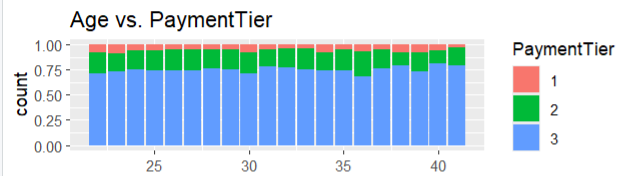
**2.** **plot2:** Compară distribuția nivelului de salarizare în funcție de anul în care angajații s-au alăturat organizației.

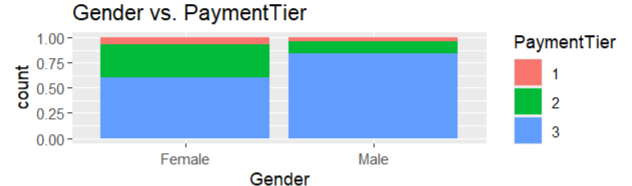


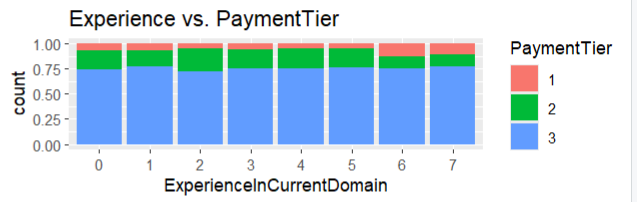
**3. plot3:** Arată distribuția nivelului de salarizare în diferite orașe.



**4. plot4:** Prezintă distribuția nivelului de salarizare în funcție de vârsta angajaților.



**5.** **plot5:** Explorează distribuția nivelului de salarizare în funcție de genul angajaților.  


**6. plot6:** Ilustrează distribuția nivelului de salarizare în raport cu experiența angajaților în domeniul lor actual.  


**7. Ploturi tip ‘Pie Chart’**

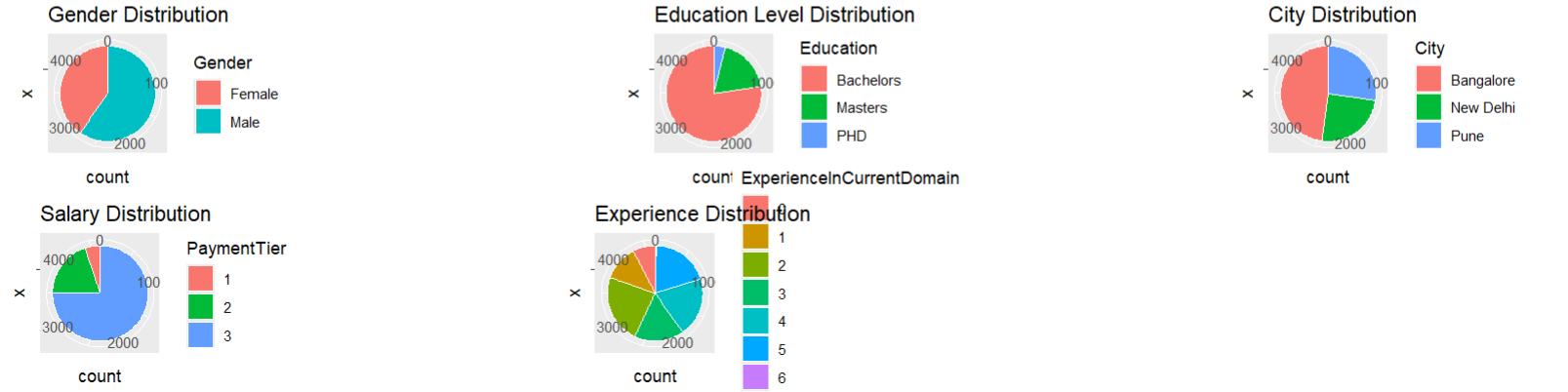
**plot7:** Reprezintă distribuția genului angajaților sub formă de diagramă polară.

**plot8:** Afișează distribuția nivelului de educație sub formă de diagramă polară.

**plot9:** Arată distribuția orașelor de reședință a angajaților sub formă de diagramă polară.

**plot10:** Prezintă distribuția nivelului de salarizare sub formă de diagramă polară.

**plot11:** Explorează distribuția experienței angajaților în domeniul lor actual sub formă de diagramă polară.



Linia **grid.arrange** organizează aceste grafice într-o grilă cu 3 coloane pentru a le afișa într-o singură vizualizare. Aceste vizualizări vor ajuta la înțelegerea relațiilor dintre caracteristicile angajaților și nivelul lor de salarizare.

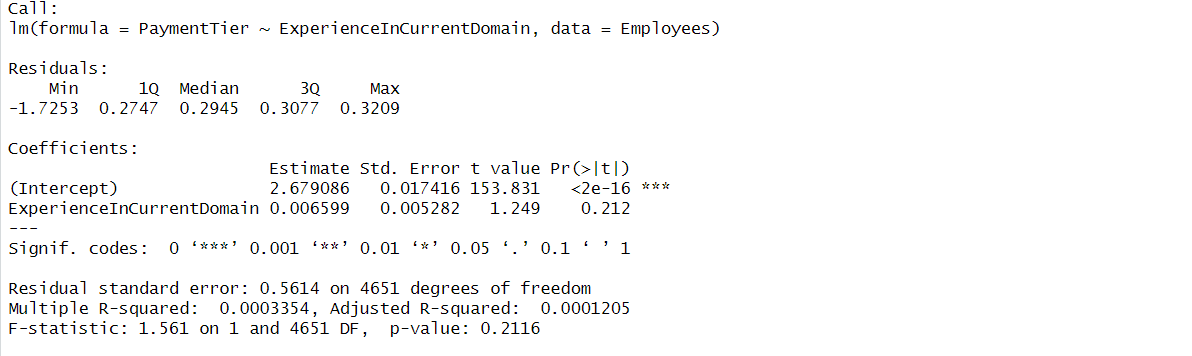
Analiza Impactului Experienței Profesionale asupra Salarizării

Prin aplicarea regresiei liniare între variabila dependentă PaymentTier (nivelul de salarizare) și variabila independentă ExperienceInCurrentDomain (experiența în domeniul actual), încercăm să înțelegem cât de mult influențează experiența profesională a angajaților asupra nivelului lor de salarizare. Alegerea acestor două atribute este justificată de faptul că experiența în domeniu este unul dintre factorii importanți luați în considerare în procesul de stabilire a salariilor în multe organizații.

Iată de ce am luat în considerare aceste două atribute:

**PaymentTier (Nivelul de Salarizare):** Această variabilă reprezintă nivelul de salarizare al angajaților. Este esențială pentru că este variabila pe care dorim să o analizăm și să o prezicem în funcție de alte caracteristici ale angajaților. Înțelegerea modului în care alte variabile influențează nivelul de salarizare este crucială pentru a optimiza procesele de recrutare, retenție și dezvoltare a personalului.

**ExperienceInCurrentDomain (Experiența în Domeniul Actual):** Experiența profesională a angajaților în domeniul lor actual este un factor-cheie care poate afecta nivelul lor de salarizare. Oamenii cu o experiență mai mare într-un anumit domeniu pot fi adesea remunerați mai bine decât cei cu mai puțină experiență. Prin urmare, luarea în considerare a acestei variabile ne permite să investigăm cât de mult contribuie experiența în domeniul actual la determinarea nivelului de salarizare al angajaților.



**Reziduurile:**

Reziduurile reprezintă diferența dintre valorile observate ale variabilei dependente și valorile estimate de model.

În cazul nostru, reziduurile variază de la -1.7253 la 0.3209.

Aceste valori arată cât de departe sunt valorile observate de valorile estimate de model. Valorile mai apropiate de zero indică o mai bună potrivire a modelului.

**Coeficienții:**

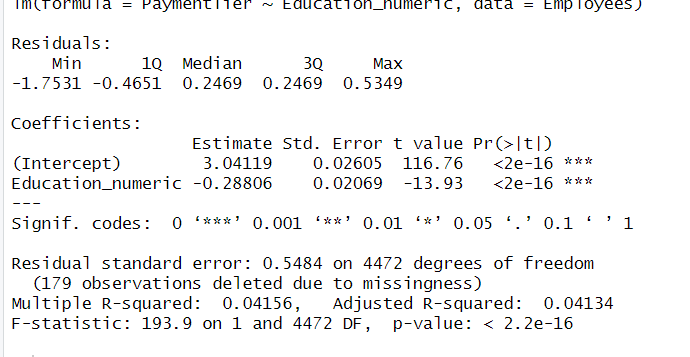
Pentru intercept (Intercept), coeficientul este 2.679086, cu o eroare standard de 0.017416. Acest lucru înseamnă că atunci când toate celelalte variabile independente sunt zero, valoarea medie a variabilei dependente este de aproximativ 2.68.

Pentru ExperienceInCurrentDomain, coeficientul este 0.006599, cu o eroare standard de 0.005282. Acest coeficient indică că pentru fiecare unitate de creștere în experiența în domeniul curent, valoarea medie a variabilei dependente crește cu aproximativ 0.0066, dar nu este semnificativ diferită de zero la un nivel obișnuit de semnificație de 0.05 (p-value = 0.212).

**Concluzie:**

Rezultatele sugerează că interceptul este semnificativ diferit de zero, iar experiența în domeniul curent nu este un predictor semnificativ al nivelului de plată, cel puțin în cadrul acestui model și set de date.

Relația dintre Nivelul de Educație și Nivelul de Plată în rândul Angajaților cu o variabila ordinala



Residuals:

Reziduurile reprezintă diferența dintre valorile observate ale variabilei răspuns (nivelul real de plată al angajaților în acest caz) și valorile prezise de model.

Minimul, prima pătrime (1Q), medianul, a treia pătrime (3Q) și maximul reziduurilor (-1.7531, -0.4651, 0.2469, 0.2469, 0.5349) indică variația acestora în setul de date.

Aceste valori furnizează informații despre distribuția reziduurilor în jurul liniei de regresie. De exemplu, medianul (0.2469) arată că reziduurile sunt în general mici și relativ simetrice în jurul valorii zero, indicând că modelul se potrivește bine datelor în medie.

Intercept:

Interceptul (3.04119) reprezintă estimarea valorii medii a nivelului de plată pentru angajații cu nivelul de educație de referință (în acest caz, probabil cei cu licență) atunci când celelalte variabile din model sunt zero sau la nivelul lor de referință.

**Nivelul de educație numeric (Education\_numeric):**

Coeficientul pentru nivelul de educație numeric (-0.28806) indică modificarea medie a nivelului de plată asociată cu o unitate de creștere a nivelului de educație numeric (adică o trecere de la un nivel de educație la altul).

În acest model, nivelul de educație numeric este construit astfel încât să crească cu fiecare nivel de educație, de la 1 pentru licență la 2 pentru masterat și 3 pentru doctorat. Astfel, un coeficient negativ sugerează că nivelurile de plată sunt, în medie, mai mici pentru angajații cu un nivel de educație mai mare (de exemplu, masterat sau doctorat) în comparație cu cei cu licență, atunci când celelalte variabile sunt constante.

P-value-ul mic (<2e-16) indică faptul că coeficientul pentru nivelul de educație numeric este semnificativ diferit de zero, ceea ce înseamnă că există suficiente dovezi pentru a sugera că există o relație semnificativă între nivelul de educație și nivelul de plată.

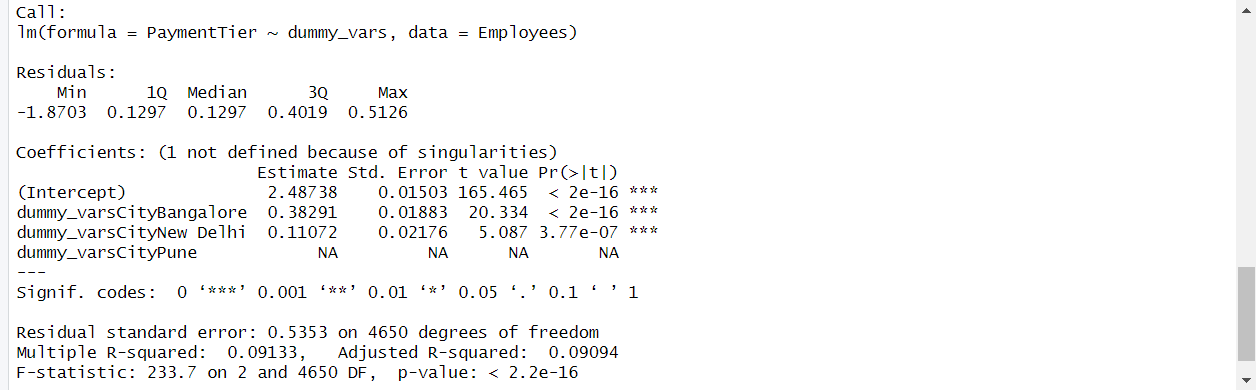
**Residual standard error (Eroarea standard a reziduurilor):**

Eroarea standard a reziduurilor (0.5484) este o măsură a dispersiei sau variabilității reziduurilor în jurul valorilor estimate ale nivelului de plată. Cu cât această valoare este mai mică, cu atât modelul este mai precis în estimarea nivelului de plată.

**Concluzie:**

In acest model, nivelul de educație are o influență semnificativă asupra nivelului de plată, cu toate că coeficientul este negativ, indicând că angajații cu niveluri de educație mai mari (masterat sau doctorat) tind să aibă niveluri de plată mai mici, în medie, decât cei cu licență, atunci când celelalte variabile sunt constant

Analiza Impactului Orașului Asupra Nivelului de Plată în Organizație: O Perspectivă prin Regresie Liniară



**Residuals:** Acestea sunt diferențele dintre valorile observate ale variabilei dependente (nivelul de plată) și valorile estimate de modelul de regresie. Valorile minime și maxime reprezintă cel mai mic și cel mai mare decalaj între valorile observate și cele prezise, iar cele de la 1Q la 3Q reprezintă cuartile 1 până la 3 ale acestor decalaje.

**Coefficients:** Coeficienții de regresie indică schimbarea medie a variabilei dependente pentru fiecare unitate de schimbare a variabilei independente, menținând celelalte variabile independente constante. În acest caz, interceptul este valoarea estimată a nivelului de plată pentru angajații din orașul Pune. Coeficienții pentru dummy\_vars reflectă diferența estimată în nivelul de plată între fiecare oraș și orașul Pune (care este variabila de referință). Este important de remarcat că coeficientul pentru dummy\_varsCityPune este marcat cu "NA", indicând că acesta nu este definit din cauza singularității (probabil din cauza utilizării unui oraș ca referință).

**Interpretare:**

Intercept: Interceptul de 2.48738 reprezintă nivelul de plată mediu pentru angajații din orașul Pune, care este orașul de referință în acest model. Acesta este nivelul de plată când toate celelalte variabile independente sunt zero.

**Coeficientul pentru dummy\_varsCityBangalore**: Acesta este 0.38291, ceea ce înseamnă că angajații din Bangalore au în medie un nivel de plată cu 0.38291 unități mai mare decât cei din Pune, ținând cont de celelalte variabile constante.

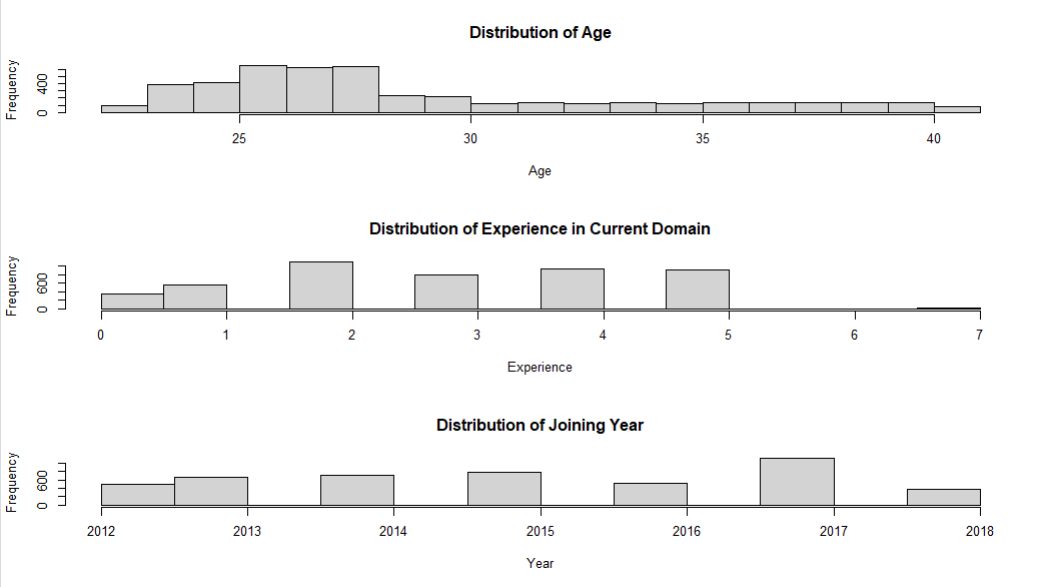
**Coeficientul pentru dummy\_varsCityNew Delhi:** Acesta este 0.11072, indicând că angajații din New Delhi au un nivel de plată mediu cu 0.11072 unități mai mare decât cei din Pune, ținând cont de celelalte variabile constante.

**dummy\_varsCityPune:** Deși coeficientul este listat ca "NA", deoarece acesta este orașul de referință, în interpretare, acesta ar fi 0, deoarece este diferența dintre nivelul de plată al angajaților din Pune și ei înșiși, ceea ce este evident zero.

**Concluzie:**

Singularitatea în coeficientul pentru orașul Pune indică o problemă potențială, deoarece Pune este ales ca oraș de referință. Pentru a evita această problemă, ar trebui să se reconsidere alegerea orașului de referință sau să se utilizeze altă metodă de codificare pentru variabilele categorice.

**Vizualizarea distribuției datelor demografice:**

****

**Prezicerea nivelului de plată pentru angajați folosind regresia liniară**

Educație: Bachelors

Anul de angajare: 2015

Orașul: Pune

Vârsta: 30 de ani

Genul: Masculin

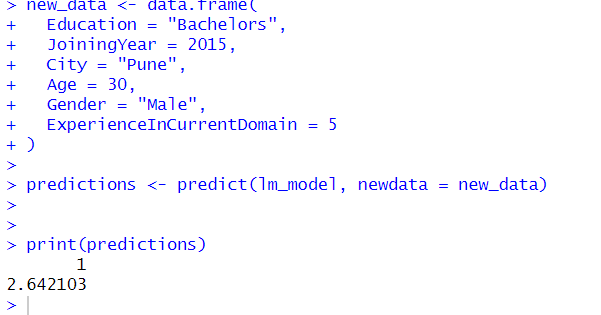
Experiența în domeniu: 5 ani

În cadrul modelului de regresie liniară, coeficienții asociate fiecărei variabile independent explică modul în care aceste variabile contribuie la variabilitatea variabilei dependente (la noi, nivelul de plata)

Pentru fiecare unitate de creștere a unei variabile independente (spre exemplu, vârsta), coeficientul corespunzător indică cât de mult este de așteptat să crească sau să scadă nivelul de plată al angajatului. În cazul nostru, deoarece vârsta este o variabilă numerică, coeficientul asociat vârstei arată schimbarea medie a nivelului de plată pentru fiecare an de creștere a vârstei.

Pentru a interpreta această valoare, trebuie să ne uităm la fiecare coeficient din modelul de regresie și la valorile corespunzătoare din setul de date. De exemplu, dacă coeficientul pentru vârstă este pozitiv și semnificativ, acest lucru ar indica faptul că o creștere a vârstei este asociată cu o creștere a nivelului de plată. În cazul în care coeficientul este negativ și semnificativ, ar însemna că o creștere a vârstei este asociată cu o scădere a nivelului de plată.

In cazul de fata, am folosit functia **predict()** pentru a face o predicție pe baza noului set de date creat anterior și a modelului de regresie liniară antrenat (**predictions <- predict(lm\_model, newdata = new\_data**) dupa care am obtinut rezultatul **2.642103.** Valoarea de 2.642103 reprezintă estimarea nivelului de plată al unui angajat cu caracteristicile specificate și este calculată în baza coeficienților și valorilor caracteristicilor din modelul de regresie liniară.

****

Estimarea și Intervalul de Confidență pentru Coeficienții de Regresie

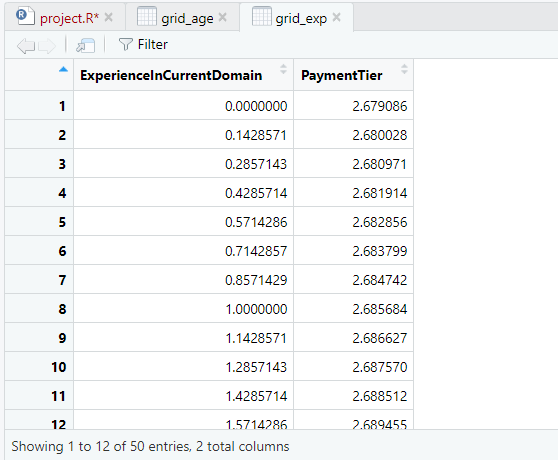
In continuare, am creat două modele de regresie liniară.

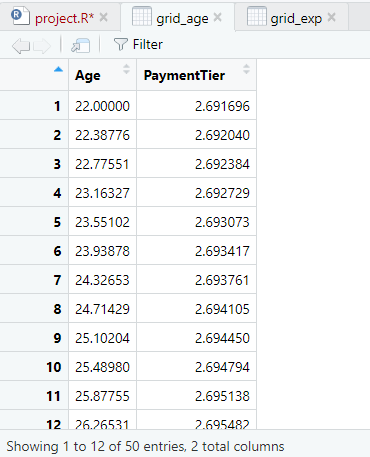
Primul model, **mod\_payment\_exp**, utilizează variabila **ExperienceInCurrentDomain** pentru a prezice **PaymentTier**, iar al doilea model, **mod\_payment\_age**, utilizează variabila **Age** în aceeași manieră.

Apoi, am definit intervalele de valori pentru **ExperienceInCurrentDomain** și **Age**, luând minimul și maximul din setul de date și creând 50 de puncte echidistante între aceste valori.

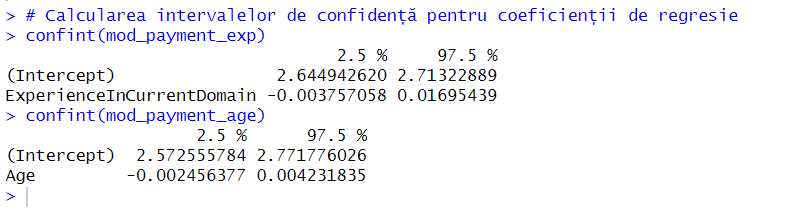
Folosind funcția predict, am estimat valorile **PaymentTier** pentru fiecare punct din intervalul **ExperienceInCurrentDomain** și **Age**, bazându-ne pe modelele de regresie create anterior.

În cele din urmă, am afișat aceste estimări în cadre de date separate pentru fiecare variabilă independentă, **ExperienceInCurrentDomain** și **Age**, sub numele de grid\_exp și grid\_age. Aceste cadre de date conțin estimările PaymentTier pentru fiecare valoare din intervalele respective.





Pentru a evalua incertitudinea asociată cu coeficienții de regresie, am calculat, de asemenea, intervalele de confidență folosind funcția confint pentru ambele modele. Aceste intervale oferă o gamă de valori în care coeficienții de regresie sunt susceptibili să cadă, cu o anumită probabilitate.



**Pentru modelul mod\_payment\_exp:**

-Coeficientul pentru intercept ((Intercept)) are un interval de confidență de la 2.644942620 la 2.71322889.

-Coeficientul pentru variabila independentă ExperienceInCurrentDomain are un interval de confidență de la -0.003757058 la 0.01695439.

**Pentru modelul mod\_payment\_age:**

-Coeficientul pentru intercept ((Intercept)) are un interval de confidență de la 2.572555784 la 2.771776026.

-Coeficientul pentru variabila independentă Age are un interval de confidență de la -0.002456377 la 0.004231835.

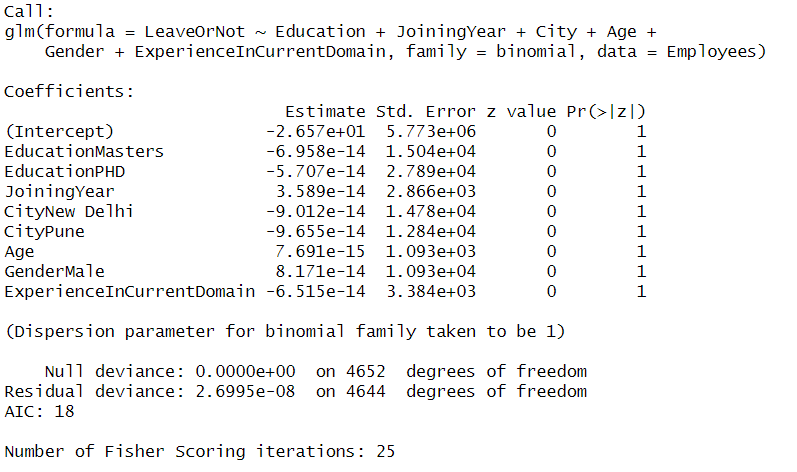
Interpretarea acestor intervale de confidență este că avem un grad rezonabil de siguranță că valorile reale ale coeficienților de regresie sunt incluse în aceste intervale. De exemplu, pentru modelul mod\_payment\_exp, avem un interval de confidență pentru coeficientul ExperienceInCurrentDomain care include zero. Acest lucru sugerează că există o incertitudine considerabilă în ceea ce privește efectul variabilei ExperienceInCurrentDomain asupra variabilei răspuns PaymentTier. Similar, pentru modelul mod\_payment\_age, intervalul de confidență pentru coeficientul Age include și zero, indicând că există incertitudine în efectul vârstei asupra nivelului de plată.

Analiza probabilității de plecare a angajaților folosind regresia logistică

În continuare, vom analiza probabilitatea de plecare a angajaților folosind regresia logistică. Pentru a începe, trebuie să ne asigurăm că datele noastre sunt pregătite pentru analiză. Așadar, vom transforma variabila **LeaveOrNot** într-o variabilă numerică și vom recoda valorile pentru a fi compatibile cu analiza de regresie logistică.

Următorul pas este construirea modelului nostru. Vom utiliza funcția **glm()** pentru a ajusta modelul nostru de regresie logistică. Variabila dependentă va fi **LeaveOrNot**, iar variabilele independente vor fi Education, JoiningYear, City, Age, Gender și ExperienceInCurrentDomain.

După ce am construit modelul, vom examina rezultatele utilizând funcția **summary()**. Acest lucru ne va oferi coeficienții estimatăi, valorile p și alte metrici relevante pentru evaluarea performanței modelului. Astfel, vom putea să înțelegem mai bine influența fiecărei variabile asupra probabilității de plecare a angajaților.



Aceste rezultate sunt obținute dintr-un model de regresie logistică și sunt interpretate în raport cu log-odds-ul de bază al probabilității de plecare a angajaților.

**Intercept:** Coeficientul intercept (Intercept) este asociat cu log-odds-ul de bază al plecării angajaților. Este important să observăm că acest coeficient are o valoare negativă masivă, ceea ce sugerează un log-odds de bază foarte scăzut al plecării angajaților. Totuși, valoarea ridicată a p-ului indică lipsa de semnificație statistică a acestui coeficient.

Variabilele independente (Education, JoiningYear, City, Age, Gender, ExperienceInCurrentDomain): Coeficienții pentru fiecare variabilă independentă indică modificările asociate cu log-odds-ul de bază pentru fiecare unitate de schimbare în variabila respectivă, ținând celelalte constante. De exemplu, un coeficient pozitiv pentru JoiningYear indică o creștere a log-odds-ului de bază al plecării pe măsură ce anul de angajare crește. Totuși, valorile p ridicate pentru toți acești coeficienți indică lipsa de semnificație statistică a acestora, sugerând că nu există o asociere semnificativă între aceste variabile și probabilitatea de plecare a angajaților.

**Devianța (Deviance):** Devianța este o măsură a adecvării modelului la date. Valorile reduse de devianță (Residual deviance) indică o bună potrivire a modelului datelor.

Rezultatele modelului de regresie logistică arată că devianța este foarte scăzută, iar coeficienții nu sunt semnificativi statistic, ceea ce sugerează că modelul poate să nu fie foarte adecvat pentru a prezice probabilitatea de plecare a angajaților. Această discrepanță poate fi interpretată în felul următor: deși modelul pare să se potrivească bine datelor (datorită devianței reduse), coeficienții estimativi nu sunt semnificativi statistic, ceea ce indică faptul că variabilele independente incluse în model poate că nu sunt suficient de relevante sau că nu există o relație liniară clară între acestea și variabila dependentă.