Test technique Quantmetry

Vincent Gigliobianco

# 1. Statistiques descriptives

## 1. Description du jeu de données

Dans un premier temps, l’identification des valeurs manquantes et de données aberrantes a été réalisée dont le cas des candidats ayant des notes élevées pour un obtenir un dataset nettoyé.

Dans mes analyses, j’aborde la description de la variable cible embauche, les analyses univariées et bivariées des variables ainsi qu’une partie transformat ion des données pour compléter l’analyse en vue de d’identifier des variables à sélectionner pour la partie Machine Learning.

Format des colonnes: ‘data.frame’: 20000 obs. of 11 variables: $ date : Factor w/ 1827 levels "“,”2010-01-01“,..: 885 477 982 548 951 1505 1412 801 1752 521 … $ cheveux : Factor w/ 5 levels”“,”blond“,”brun“,..: 5 2 2 3 5 4 3 5 4 4 … $ age : num 25 35 29 NA 35 37 33 31 43 28 … $ exp : num 9 13 13 12 6 8 12 10 10 11 … $ salaire : num 26803 38166 35207 32442 28533 … $ sexe : Factor w/ 3 levels”“,”F“,”M“: 2 3 3 3 2 3 3 3 3 3 … $ diplome : Factor w/ 5 levels”“,”bac“,”doctorat“,..: 4 4 4 4 4 5 5 4 3 5 … $ specialite: Factor w/ 5 levels”“,”archeologie“,..: 5 4 5 5 3 5 5 4 5 4 … $ note : num 97.1 63.9 78.5 45.1 81.9 … $ dispo : Factor w/ 3 levels”“,”non“,”oui": 2 2 2 2 2 2 3 3 2 2 … $ embauche : int 0 0 0 0 0 1 0 0 1 1 …

### 1. Valeurs manquantes et aberrantes

Après import des données, les variables continues dont les distributions sont normales, permettent d’identifier rapidement des données non adaptées au cas métier telles que des candidats très jeunes ou très vieux, des notes à l’exercice strictement supérieures à 100. Des analyses univariées indiquent un pourcentage proche de 0,5% de valeurs à NA pour chacune des variables age, exp, salaire et note. Elles ont aussi indiqué un pourcentage proche de 0,5% de valeurs à vide pour chacune des variables catégorielles date, cheveux, sexe, diplome, specialite et dispo.

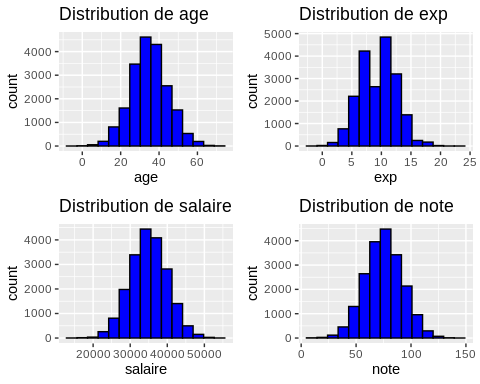


Table continues below

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| -2 | -1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 1 | 23 | 45 | 111 | 282 | 483 | 871 | 1341 | 1878 | 2346 | 2638 | 2475 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 23 |
| 2371 | 1849 | 1357 | 887 | 501 | 247 | 127 | 48 | 16 | 4 | 1 | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
| age | exp | salaire | note |
| 91 | 96 | 95 | 114 |

## *0*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | blond | brun | chatain | roux | NA |
| 103 | 5929 | 5991 | 6022 | 1955 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | F | M | NA |
| 100 | 8011 | 11889 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | bac | doctorat | licence | master | NA |
| 110 | 2453 | 2547 | 7377 | 7513 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | archeologie | detective | forage | geologie | NA |
| 93 | 1384 | 4155 | 4340 | 10028 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| date | cheveux | sexe | diplome | specialite | dispo |
| 91 | 103 | 100 | 110 | 93 | 106 |

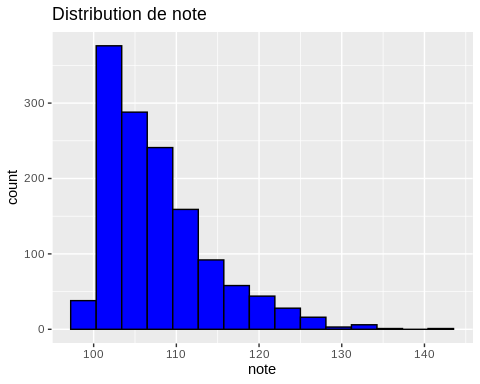
Ainsi pour nettoyer le dataset j’ai donc envisagé:

* de restreindre l’âge des candidats entre [17;65]
* Pour l’expérience, deux valeurs négatives ont été ôtées
* Les valeurs manquantes identifiées ont été ôtées pour les variables date, cheveux, sexe, diplome, specialite et dispo ainsi que pour les variables age, exp, salaire
* Enfin, pour le cas des candidats ayant eu des notes supérieures à 100, la distribution dissymétrique vers la gauche a montré que la majorité des candidats possède une note assez proche de 100. J’ai donc décidé de garder des observations.

Ces actions ont permis d’obtenir 18320 observations.

|  |  |  |  |
| --- | --- | --- | --- |
| archeologie | detective | forage | geologie |
| 188 | 386 | 322 | 455 |

|  |  |  |  |
| --- | --- | --- | --- |
| bac | doctorat | licence | master |
| 25 | 653 | 135 | 538 |



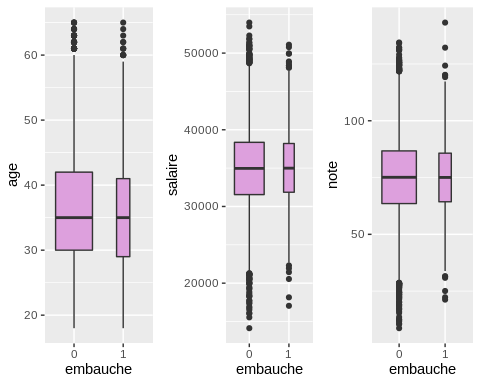
### 2. Profil de la variable cible embauche

La distribution de la variable cible embauche ci-dessous nous indique qu’il y a 2090/( 16230 + 2090) = 11,41 % d’embauches

La variable “target” embauche est donc assez déséquilibrée, mais on considère que ce déséquilibre n’est pas trop grave pour notre étude. On pourra néanmoins pour la partie Machine Learning envisager une métrique de mesure des performances plus adaptée qu’une simple mesure du type “accuracy”

### 4. Liaisons des variables continues et catégorielles et liaisons avec la variable embauche

|  |  |  |  |
| --- | --- | --- | --- |
|  | age | salaire | note |
| **age** | 1 | 0.002711 | -0.001995 |
| **salaire** | 0.002711 | 1 | -0.4506 |
| **note** | -0.001995 | -0.4506 | 1 |



Kruskal-Wallis rank sum test: age by embauche

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 3.765 | 1 | 0.05233 |

* **method**: Wilcoxon rank sum test
* **data.name**: df\_filter\_nomissingembauche
* **p.value**:

|  |
| --- |
| * 0 |
| * 0.05233 |

* **p.adjust.method**: BH

Kruskal-Wallis rank sum test: note by embauche

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 0.0247 | 1 | 0.8751 |

* **method**: Wilcoxon rank sum test
* **data.name**: df\_filter\_nomissingembauche
* **p.value**:

|  |
| --- |
| * 0 |
| * 0.8751 |

* **p.adjust.method**: BH

Kruskal-Wallis rank sum test: salaire by embauche

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 0.3296 | 1 | 0.5659 |

* **method**: Wilcoxon rank sum test
* **data.name**: df\_filter\_nomissingembauche
* **p.value**:

|  |
| --- |
| * 0 |
| * 0.5659 |

* **p.adjust.method**: BH

Pearson’s Chi-squared test: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 1.665 | 3 | 0.6446 |

Pearson’s Chi-squared test with Yates’ continuity correction: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 14.65 | 1 | 0.0001297 \* \* \* |

Pearson’s Chi-squared test: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 51.68 | 3 | 3.513e-11 \* \* \* |

Pearson’s Chi-squared test: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 122.2 | 3 | 2.548e-26 \* \* \* |

Pearson’s Chi-squared test with Yates’ continuity correction: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 2.444 | 1 | 0.118 |

Une analyse de la matrice des corrélations a mis en évidence que seule la variable salaire est un peu corrélée négativement, avec la note à l’exercice (corrélation = -0.45). Il est peut être pertinent de penser que la note à l’exercice peut être un facteur d’explication et de la prédiction du succès ou de l’échec d’une candidature.

Les boxplots ainsi que des tests de Kruskall Wallis non significatifs indiquent que les variables age, salaire et note prises toutes seules n’ont pas de lien fort avec la variable embauche. On peut penser aussi, que le nombre d’années d’expérience et le fait d’être spécialiste dans un domaine sont des candidats possibles à l’explication et de la prédiction du succès ou de l’échec d’une candidature.

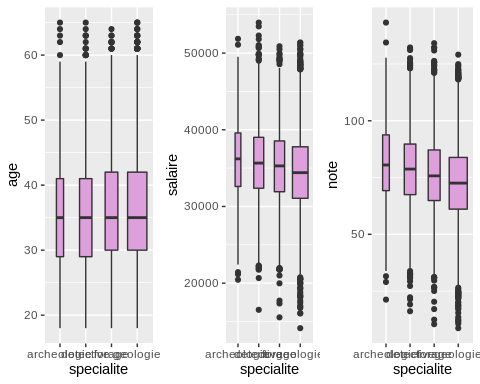
J’ai analysé les liaisons entre embauche et les variables catégorielles cheveux, sexe, diplome, specialite et dispo en sortant tous les statistiques du Chi2 entre ces variables. Et l’analyse de la variable nombre d’années d’expérience est proposée plus tard dans le document en raison de la nature ordinale de exp et de son grand nombre de modalités.

J’ai constaté qu’il y a des liaisons significatives pour specialite, diplome, et sexe. Ces variables catégorielles sont donc celles qui vont en principe expliquer le mieux la variable embauche. Ceci indique que les pourcentages de candidats embauchés différent significativement pour chacune des modalités de specialite, diplome et sexe.

Le tableau croisé entre specialite et embauche pour lequel on note un pourcentage de 20% de candidats embauchés pour les archéologues.

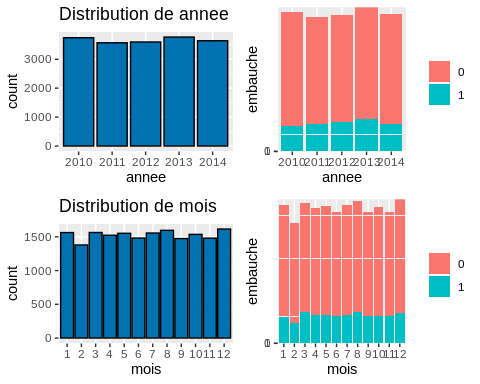
|  |  |  |  |
| --- | --- | --- | --- |
| specialite | embauche 0 | 1 | Total |
| **archeologie** N Row(%) Column(%) | 1015 79.9% 6.3% | 256 20.1% 12.2% | 1271 6.9% |
| **detective** N Row(%) Column(%) | 3411 90.0% 21.0% | 378 10.0% 18.1% | 3789 20.7% |
| **forage** N Row(%) Column(%) | 3641 90.7% 22.4% | 373 9.3% 17.8% | 4014 21.9% |
| **geologie** N Row(%) Column(%) | 8163 88.3% 50.3% | 1083 11.7% 51.8% | 9246 50.5% |
| Total | 16230 88.6% | 2090 11.4% | 18320 |

Plusieurs croisements sont possibles mais je propose ici un résultat intéressant entre les variables age, salaire et note et la variable specialite: on constate que les notes, les salaires et les âges varient beaucoup entre les archéologues, les détectives, les spécialistes en forage et les géologues.



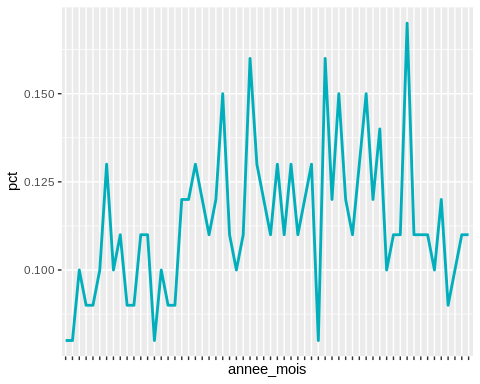
#### 5. Analyse de la variable date de candidature

On peut se demander si cette date de candidature n’influe pas sur l’embauche. Après création des variables “annee” et “mois” et analyse (graphique) de la liaison entre la variable mois et la variable embauche, le mois tel qu’il est codé avec ses 12 modalités est apparu sans lien statistique avec l’embauche. Mais j’ai voulu analyser l’évolution du % d’embauchés pour chaque mois écoulé.



Pearson’s Chi-squared test: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 11.92 | 11 | 0.3694 |



Tout en notant le caractère très irrégulier du % de candidats embauchés, on peut s’interroger sur la possibilté de créer une feature qui tienne compte d’une période pour laquelle le % d’embauche est plutôt élevée ou plutôt faible.

Par exemple, on pourrait créer:

* une feature “année” + “mois” en vue de tester cette variable lors de la phase de Machine Learning, mais le nombre de modalités devraient être diminué au préalable.
* En lien avec les pratiques de features engineering, une feature moyenne des % d’embauche ou une feature calculée grâce à la moyenne des % d’embauchés durant 15 jours précédant une date de candidature.

#### 6. Intérêt de la transformation des variables avec quartiles et déciles

Je choisis de recoder les variables continues age, note, salaire et aussi la variable ordinale nombre d’années d’expérience grâce à des quartiles pour obtenir de nouvelles variables possédant des effectifs équilibrés par modalité. Des quartiles permettent de faire de l’analyse exploratoire pour identifier éventuellement, si ces variables recodées peuvent avoir un lien statistique avec la variable cible embauche et aussi surtout à pouvoir identifier des interactions entre ces nouvelles variables et les autres comme specialite, diplome, sexe, cheveux et dispo.

Un autre recodage en déciles peut aussi servir en vue de faire du feature engineering, notamment en utilisant un decision tree de type CHAID, afin de créer des variables optimales en termes d’explication par rapport à la variable cible de rendre possible des interactions avec les autres variables catégorielles.

Les “Chi-deux” ont montré des liaisons significatives entre les variables note\_Q, salaire\_Q et embauche.

Pearson’s Chi-squared test: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 7.591 | 3 | 0.05527 |

Pearson’s Chi-squared test: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 0.3648 | 3 | 0.9474 |

Pearson’s Chi-squared test: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 9.086 | 3 | 0.02816 \* |

Pearson’s Chi-squared test: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 8.82 | 3 | 0.03178 \* |

De même, les variables déciles de l’âge, de la note et du salaire sont apparues liées à la variable cible embauche.

Pearson’s Chi-squared test: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 17.49 | 9 | 0.04151 \* |

Pearson’s Chi-squared test: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 18.25 | 9 | 0.03234 \* |

Pearson’s Chi-squared test: mytable

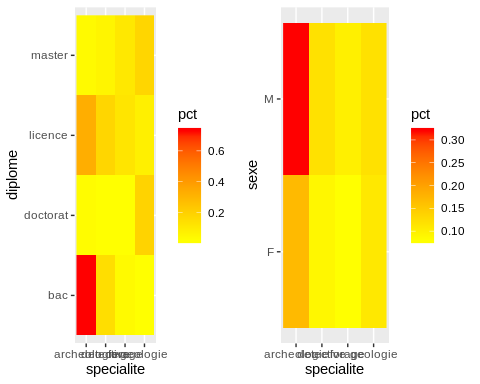
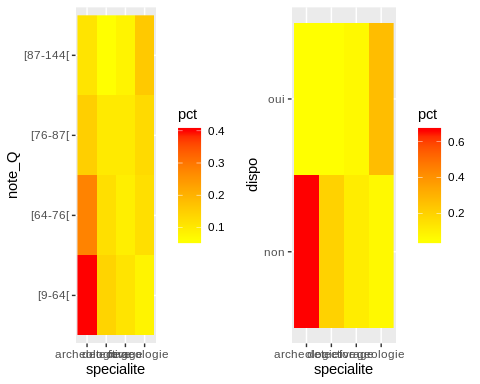
|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 37.49 | 9 | 2.153e-05 \* \* \* |

#### 7. Croisements 3D entre variables et embauche

Voici ce qu’on obtient visuellement:

* si on croise la spécialité, la note (en quartiles) et le % de candidats embauchés
* Si on croise la spécialité, la disponibilité des candidats et le % de candidats embauchés
* Si on croise la spécialité, le diplôme des candidats et le % de candidats embauchés
* Si on croise la spécialité, la sexe des candidats et le % de candidats embauchés

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| specialite | note\_Q [9-64[ | [64-76[ | [76-87[ | [87-144[ | Total |
| **archeologie** N Row(%) Column(%) | 199 15.7% 4.3% | 279 22.0% 6.1% | 299 23.5% 6.5% | 494 38.9% 10.8% | 1271 6.9% |
| **detective** N Row(%) Column(%) | 685 18.1% 15.0% | 880 23.2% 19.2% | 1029 27.2% 22.5% | 1195 31.5% 26.1% | 3789 20.7% |
| **forage** N Row(%) Column(%) | 931 23.2% 20.3% | 1011 25.2% 22.1% | 1038 25.9% 22.7% | 1034 25.8% 22.6% | 4014 21.9% |
| **geologie** N Row(%) Column(%) | 2765 29.9% 60.4% | 2415 26.1% 52.7% | 2209 23.9% 48.3% | 1857 20.1% 40.5% | 9246 50.5% |
| Total | 4580 25% | 4585 25% | 4575 25% | 4580 25% | 18320 |



Ce croisement de variables montre clairement que ce sont les archéologues qui sont les plus embauchés par rapport aux autres spécialités et ce d’autant plus que leurs notes obtenues à l’exercice ont été basses. D’autres croisements avec la spécialité ont aussi montré que : - il y a beaucoup d’embauches parmi les archéologues d’autant plus que qu’ils sont peu diplômés - il y a beaucoup d’embauches parmi les archéologues d’autant plus que qu’ils sont des hommes

### Conclusion sur les variables à sélectionner

Les résultats des Chi2 ont montré que les variables catégorielles spécialité, diplôme et sexe sont intéressantes à sélectionner en vue de la prédiction du succès ou de l’échec d’une candidature.

De plus, les statistiques du Chi2 sont significatives pour les nouvelles variables de type quartiles pour la note et le salaire. Pour ce qui concerne, les autres variables de type déciles, cette fois ce sont l’âge, la note et le salaire qui sont significatifs. Et l’expérience est à nouveau non significative pour la variable de type quartile.

Par conséquent, toutes les variables:

* spécialité, diplôme et sexe
* quartile de la note et quartile du salaire
* décile de l’âge, décile de la note et décile du salaire

sont des features intéressantes à sélectionner en vue de la prédiction du succès ou de l’échec d’une candidature. Enfin, le graphe 3D vu au paragraphe 7., indique clairement que la variable disponibilité des candidats est intéressante pour la prédiction du succès ou de l’échec d’une candidature.

## 2. Dépendances statistiques

|  |  |  |  |
| --- | --- | --- | --- |
| specialite | sexe F | M | Total |
| **archeologie** N Row(%) Column(%) | 1002 78.8% 13.6% | 269 21.2% 2.5% | 1271 6.9% |
| **detective** N Row(%) Column(%) | 2311 61.0% 31.3% | 1478 39.0% 13.5% | 3789 20.7% |
| **forage** N Row(%) Column(%) | 1826 45.5% 24.7% | 2188 54.5% 20.0% | 4014 21.9% |
| **geologie** N Row(%) Column(%) | 2244 24.3% 30.4% | 7002 75.7% 64.0% | 9246 50.5% |
| Total | 7383 40.3% | 10937 59.7% | 18320 |

Pearson’s Chi-squared test: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 2491 | 3 | 0 \* \* \* |

Table continues below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| cheveux | salaire\_Q [14128-31591[ | [31591-34964[ | [34964-38348[ | [38348-53977[ |
| **blond** N Row(%) Column(%) | 1452 26.6% 31.7% | 1402 25.7% 30.6% | 1317 24.1% 28.8% | 1294 23.7% 28.3% |
| **brun** N Row(%) Column(%) | 1334 24.1% 29.1% | 1398 25.3% 30.5% | 1401 25.4% 30.6% | 1392 25.2% 30.4% |
| **chatain** N Row(%) Column(%) | 1270 23.0% 27.7% | 1308 23.7% 28.6% | 1439 26.1% 31.4% | 1493 27.1% 32.6% |
| **roux** N Row(%) Column(%) | 525 28.8% 11.5% | 472 25.9% 10.3% | 423 23.2% 9.2% | 400 22.0% 8.7% |
| Total | 4581 25% | 4580 25% | 4580 25% | 4579 25% |

|  |
| --- |
| Total |
| 5465 29.8% |
| 5525 30.2% |
| 5510 30.1% |
| 1820 9.9% |
| 18320 |

Pearson’s Chi-squared test: mytable

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 58.73 | 9 | 2.357e-09 \* \* \* |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| exp\_Q | note\_Q [9-64[ | [64-76[ | [76-87[ | [87-144[ | Total |
| **[0-7[** N Row(%) Column(%) | 1122 24.2% 24.5% | 1160 25.0% 25.3% | 1161 25.1% 25.4% | 1190 25.7% 26.0% | 4633 25.3% |
| **[7-9[** N Row(%) Column(%) | 1162 25.3% 25.4% | 1115 24.3% 24.3% | 1151 25.1% 25.2% | 1156 25.2% 25.2% | 4584 25.0% |
| **[9-12[** N Row(%) Column(%) | 1547 25.1% 33.8% | 1578 25.6% 34.4% | 1511 24.5% 33.0% | 1536 24.9% 33.5% | 6172 33.7% |
| **[12-23[** N Row(%) Column(%) | 749 25.6% 16.4% | 732 25.0% 16.0% | 752 25.7% 16.4% | 698 23.8% 15.2% | 2931 16.0% |
| Total | 4580 25% | 4585 25% | 4575 25% | 4580 25% | 18320 |

Pearson’s Chi-squared test: mytable J’ai utilisé ici des tests du Chi-deux et choisi les versions de variables avec quartiles pour le salaire, le nombre d’années d’expérience et la note à l’exercice en vue de considérer des variables ayant un petit nombre de modalités.

|  |  |  |
| --- | --- | --- |
| Test statistic | df | P value |
| 7.188 | 9 | 0.6175 |

1. On constate grâce à la P-value très inférieure à 0.05 pour la statistique du Chi2, que la dépendance statistique est très significative entre spécialité et sexe d’autant plus que les différentes spécialités sont composées de proportions d’hommes et de femmes très différentes.
2. La P-value du test entre la couleur des cheveux et le salaire demandé indique qu’il y a une certaine dépendance statistique.
3. En revanche, il n’y a pas de dépendance statistique entre le nombre d’années d’expérience et la note à l’exercice.