UNIVERSIDAD DE GUANAJUATO
División De Ingenierías — Campus Irapuato-Salamanca

Práctica Intermedia

Minería de Datos

Jose Abel Vilchis Mar.

NUA: 390517

08 de mayo del 2020

Descripción de la práctica:

Hay un archivo CSV llamado survey_results_public.csv (comprimido como survey_results_public.rar) que contiene los datos de una encuesta recopilada de los usuarios de Stack OverFlow en 2019 con respecto a las siguientes diez variables (el nombre de los campos en el archivo están entre paréntesis): país (Country), nivel educativo (EdLevel), tipo de desarrollador (DevType), años de experiencia con codificación (YearsCode), salario anual en dólares estadounidenses (ConvertedComp), número promedio de horas de trabajo por semana (WorkWeekHrs), lenguaje de programación he / ella tiene experiencia con (LanguageWorkedWith), edad (Age), género (Gender) y etnia (Ethnicity). Hay datos para 88.883 usuarios.

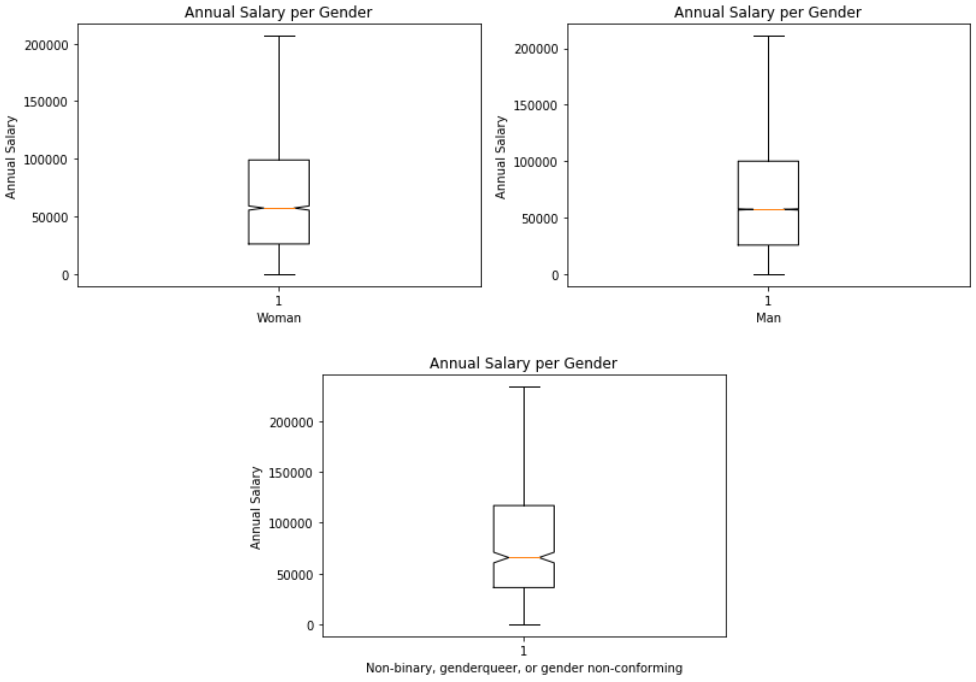
Para algunas variables, los usuarios podrían responder con más de una respuesta, con las respuestas separadas con un ; en el archivo. Por ejemplo, en el lenguaje de programación con el que tiene experiencia, un usuario puede seleccionar al mismo tiempo C; C++; JavaScript; Python. En ese caso, para las estadísticas, el mismo usuario contará por cada idioma que elija. Lo mismo se aplica para cualquier otra variable que permita múltiples respuestas.

En otros casos, los usuarios pueden omitir una o varias respuestas, y en el archivo, podemos encontrar valores NaN o valores de cadena vacíos. En ese caso, para las estadísticas, esos valores deben ignorarse.

La práctica consiste en las siguientes pequeñas tareas de procesamiento y análisis de los datos contenidos en el archivo. Para cada tarea, debe escribir una función de Python como parte del archivo de código pratice.py.

Ejercicios:

1. Calcule el resumen de cinco números, boxplot, la media y la desviación estándar para el salario anual por género.



Figuras 1, 2, 3.- Diagramas de caja del salario anual por género.

```
Woman
Min:  0.0
Max:  2000000.0
First Quartile:  26124.0
Median:  57287.0
Third Quartile:  99000.0
Mean:  114066.78108314263
Standard Desviation: 255575.45397300526
```

```
Man
Min:  0.0
Max:  2000000.0
First Quartile:  25656.0
Median:  57254.0
Third Quartile:  100000.0
Mean:  127346.65901781662
Standard Desviation: 284521.4974200705
```

```
Non-binary, genderqueer, or gender non-conforming
Min:  0.0
Max:  2000000.0
First Quartile:  36126.0
Median:  65653.0
Third Quartile:  116929.75
Mean:  154890.6462585034
Standard Desviation: 328791.4545744816
```

Figuras 4, 5, 6.- Resumen de 5 números, media y desviación estándar del salario anual por género.

En un principio los diagramas de caja de los tres datos a analizar se mantienen muy similares, difieriendo en solo unos cuantos miles. Del resumen de cinco números por medio de la media se podría diferir que en cuanto a salarios el género mejor pagado es la categoría la cual encapsula tanto las personas no binarias, *genderqueer* o no conformes con su género; seguido por el salario de los hombres, para quedar en último lugar con muy poca diferencia el de las mujeres.

No obstante, la desviación estándar de los no binarios es la mayor de entre las tres, por lo que los datos de ésta categoría están más dispersos de la media y con ello el hecho de que sea el salario más renumerado entra en duda. Por su parte, la media entre los hombres y las mujeres es muy similar, difiriendo por una cuantas centenas. Sin embargo, la desviación estándar difiere en miles, por lo que puede concluirse que los datos de las mujeres están menos dispersos con la media, y con ello es la brecha salarial entre las mujeres encuestadas es menor en comparación a los otros dos géneros.

2. Calcule el resumen de cinco números, boxplot, la media y la desviación estándar del salario anual por grupo étnico.

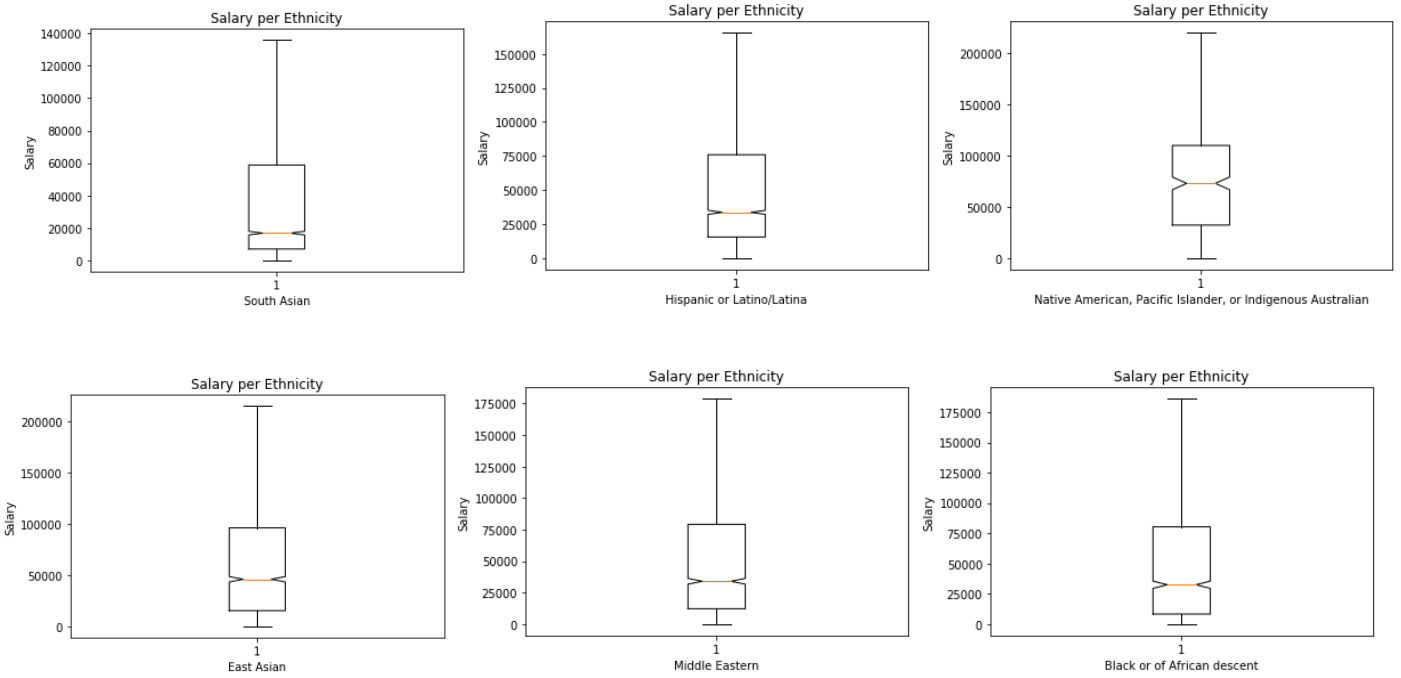
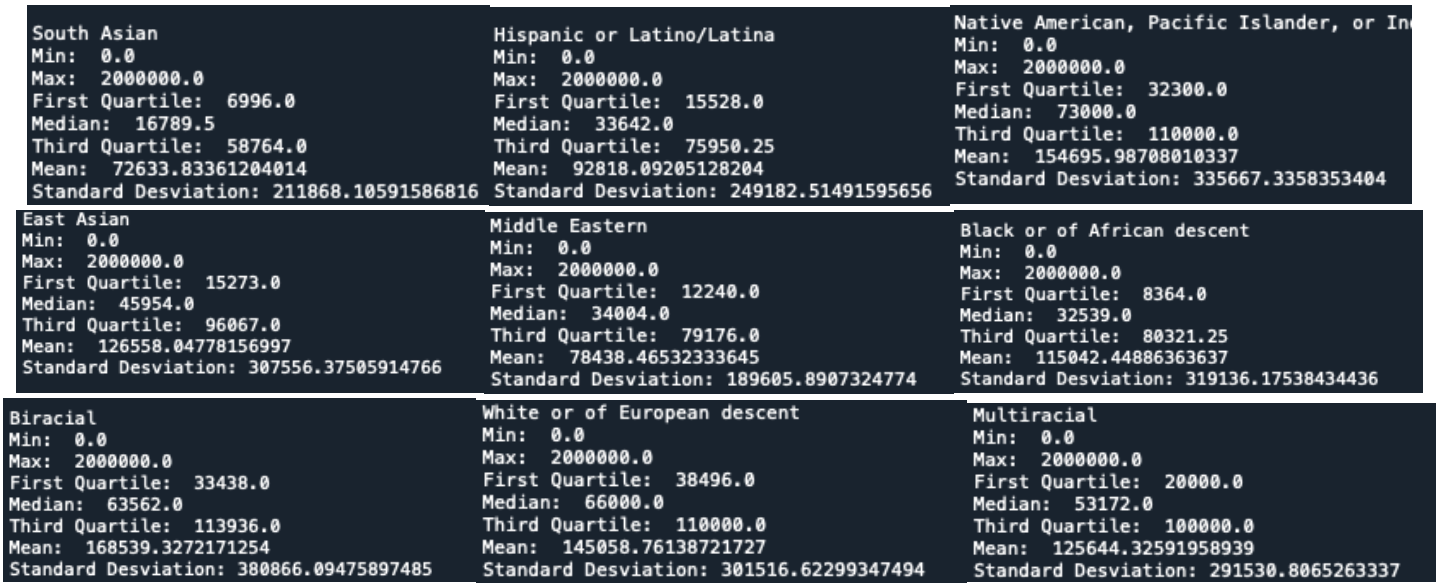


Figura 7.- Diagramas de caja del salario anual por grupo étnico.



Figuras 8 - 16.- Resumen de 5 números, media y desviación estándar del salario anual por grupo étnico.

Por medio de los diagramas de caja puede observarse que la etnia con mayor renumeración es la birracial, seguida por la que encapsula las etnias nativas de América, el Pacífico e Australia, , y quedándose como tercero la blanca o descendiente europeo. No obstante, por medio de la desviación estándar se puede observar que la de los birraciales es mucho mayor en comparación a la de los nativos y los blancos, por lo que puede inferirse que no existe una gran equidad ecónomica entre los birraciales, mientras que en los blancos dicha equidad es mayor debido a que los salarios de los encuestados están más cercanos entre sí. Dicha información puede confirmarse al observase de nuevo el diagrama de caja de los birraciles, la mitad de ellos se encuentran en un rango de entre 33438 y 64562 dolares, mientras la otra mitad está en un espectro muy ancho entre los 113936. Por lo tanto, es más factible que los mejores pagado sean lo de ascendencia blanca, y esto dice que puede que dicho grupo estén en mejores puestos trabajo.

Al analizarse los demás datos se puede observar un contraste entre las etnias de diferentes regiones el mundo, mientras que en países europeos o norteamericanos se les paga muy bien en los países del sur, y sobre todo aquellos en vías de desarrollo, suelen ser muy poco renumerados. Esto es factible debido a que no es un secreto que países como China, India, y Vietnam suelen presentar mano de obra barata, aunque América Latina de igual forma puede darse dicha observación.

3. Calcule el resumen de cinco números, boxplot, la media y la desviación estándar para el salario anual por tipo de desarrollador.

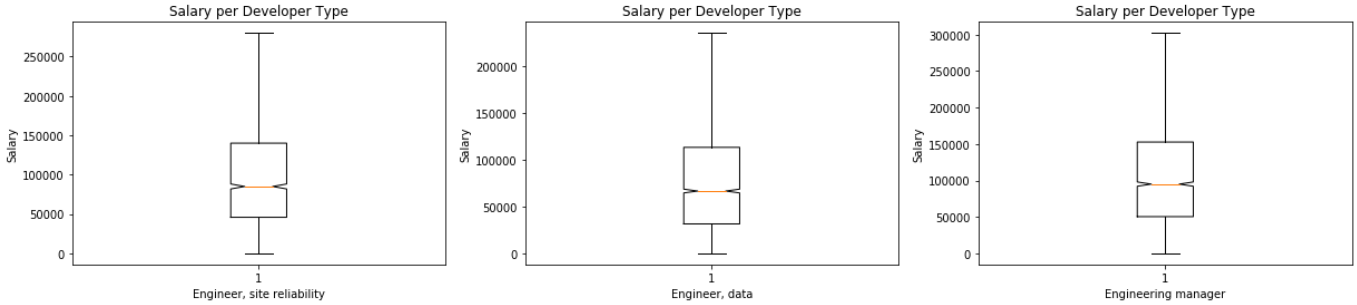
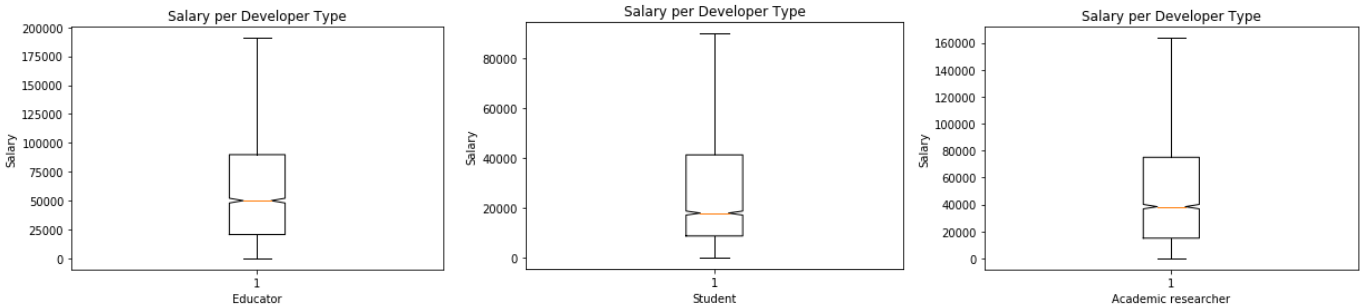


Figura 17.- Diagramas de caja del salario anual en la ingeniería.

En la primera observación se buscó ver qué diferencia había entre los distintos ingenieros, siendo el ingeniero en administración el de mayor sueldo seguido del ingeniero en confiabilidad del sitio y por último el ingeniero en datos. Su salarios son similares, difiriendo en miles, por lo que puede inferirse que aquellos que estudian una ingeniería en alguna rama de la informática presentan una equidad ecónomica más justa y estable en comparación a las demás carreras.

Figuras 18.- Diagramas de caja del salario anual en el desarrollo web.

Subsiguientemente se procedió a analizar los desarrolladores web para observar cual de los tres tipos era el mejor renumerado, siendo el desarrollador full-stack el mejor pagado, seguido por el desarrollador back-end y por últmo el desarrollador front-end. Esto nos dice que es más valorizado y buscado el hecho en que un desarrollador web tenga conocimientos en ambos campos para así no tener que contratar dos personas para realizar el trabajo que una sola podría hacer. De igual forma, el hecho de que el desarrollador back-end sea mejor renumerado nos puede decir que hay mayor demanda de dicha rama quizás a la complejidad de éste.



Figuras 19.- Diagramas de caja del salario anual en el área de la educación.

Por último se analizó el área de la educación y se encontró que es la que menor salario presenta en comparación a las otros tipos de desaarrolladores. El mejor renumerado es el educador, seguido por el investigador acádemico y por último el estudiante. En éste último puede observarse una brecha muy grande entre la primera mitad y la segunda, siendo que la mitad de los desarrolladores con nivel educativo de estudiante se concentra en un rango pequeño de entre 8664 y 16868, esto es debido a que no se posee un título por lo tanto suelen ganar menos, aunque otras variables pueden interferir como lo es la falta de experiencia o que son practicantes en alguna empresa.

4. Calcule la media, la media y la desviación estándar del salario anual por país.

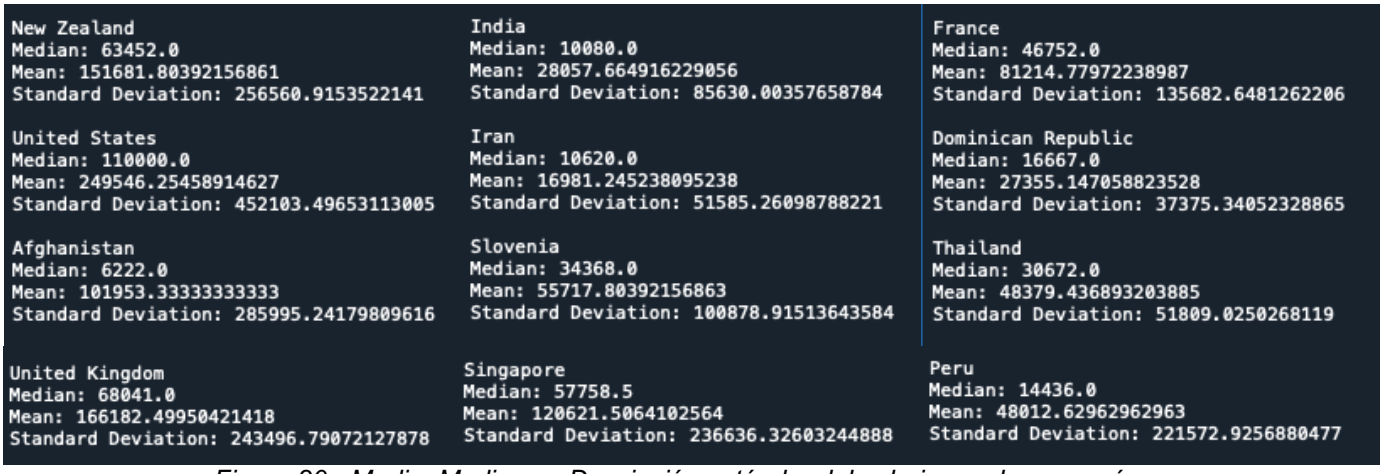


Figura 20.- Media, Mediana y Desviación estándar del salario en algunos países.

Este ejercicio está estrechamente relacionado con el ejercico número dos, solo que en este caso se orienta en países en vez de etnias.

Se omitieron algunos países donde dicha sección fue solamente respondida por una o ninguna persona. Primeramente se buscaron disntitos países pertenecientes a las regiones en vías de desarrollo, tal como lo es África, Latinoamérica, Oriente Medio, y el Sur de Asia, y se pudo observar que los salarios de dichas regiones es muy pequeño y mal renumerado, esto es debido a que algunos países en dichas regiones presentan salarios mínimos diminutos y mano de obra barata.

En contraste, los países desarrollados o llamados “primer mundo” presentaron salarios muy superiores, esto puede estar influenciado con el hecho de que la calidad de vida es mejor y por ende el salario mínimo en dichos países son mayores.

5. Obtenga un diagrama de barras con las frecuencias de respuestas para cada tipo de desarrollador.

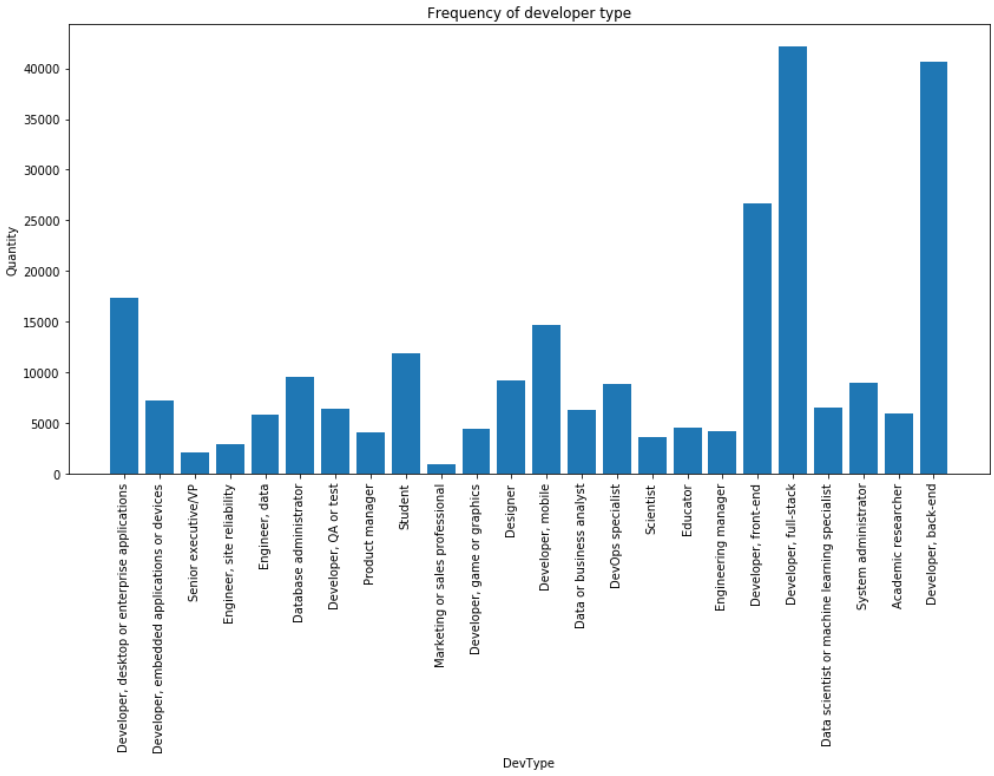


Figura 21.- Diagrama de barras de las respuesta por cada tipo de dearrollador.

Por medio de la gráfica puede observarse a simple vista que la mayoría de personas encuestadas están en el área del desarrollo software, siendo el mayor exponente el desarrollo full-stack, seguido del desarrollo back-end y front-end, aunque en sí estos tres últimos están de igual forma en el mayor exponente de la gráfica.

Por su parte, las ingeniería presentan un número de encuestados muy por debajo de la media del total de personas encuestadas, mientras que el área de la educación el mayor exponente son los estudiantes, quizás por la gran influencia que tiene le sitio web StackOverFlow a la hora de encontrar soluciones a problematicas que se tengan en el cursamiento de la carrera.

6. Trace histogramas con 10 bins para los años de experiencia con la codificación por género.

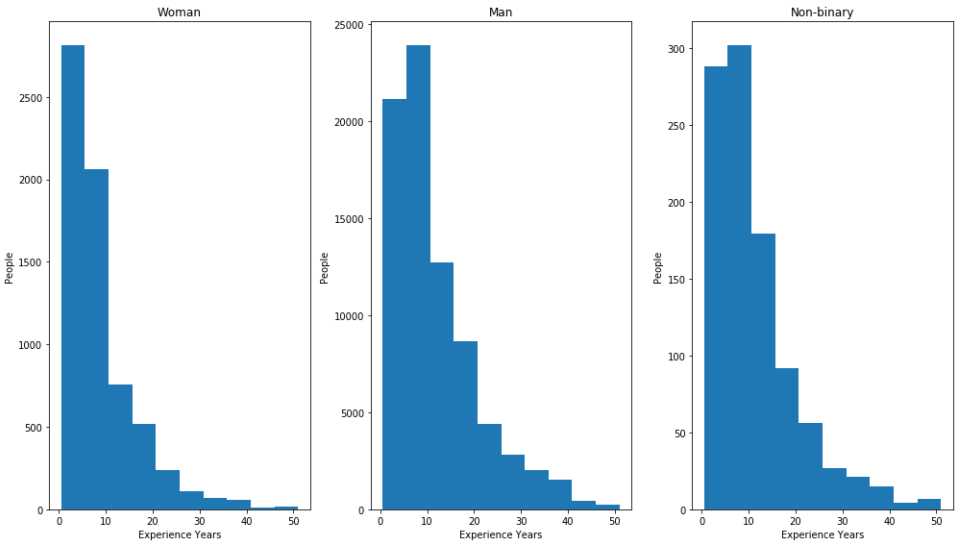


Figura 22.- Histogramas de años de experiencia por género.

Los histogramas nos dice en primera instacia que el mayor número de encuestados fueron hombres, siendo por ejemplo casi 10 veces mayor que la cantidad de mujeres. De igual forma, la forma del gráfico nos dice que los hombres suelen tener mayor número de experiencia que el de las mujeres o las personas no binarias, comprendiéndose entre 0 y 15 años la mayoría de éstos. Estos resultados pueden estar influencias con el hecho que el área del desarrollo en software suele estar mayormente dominado por los hombres, y con ello la presencia de mujeres o personas no binarias suele verse invisibilizado, aunque en los últmos años los demás géneros se han ido interesando en dicha área.

7. Trace histogramas con 10 bins para el número promedio de horas de trabajo por semana, por tipo de desarrollador.

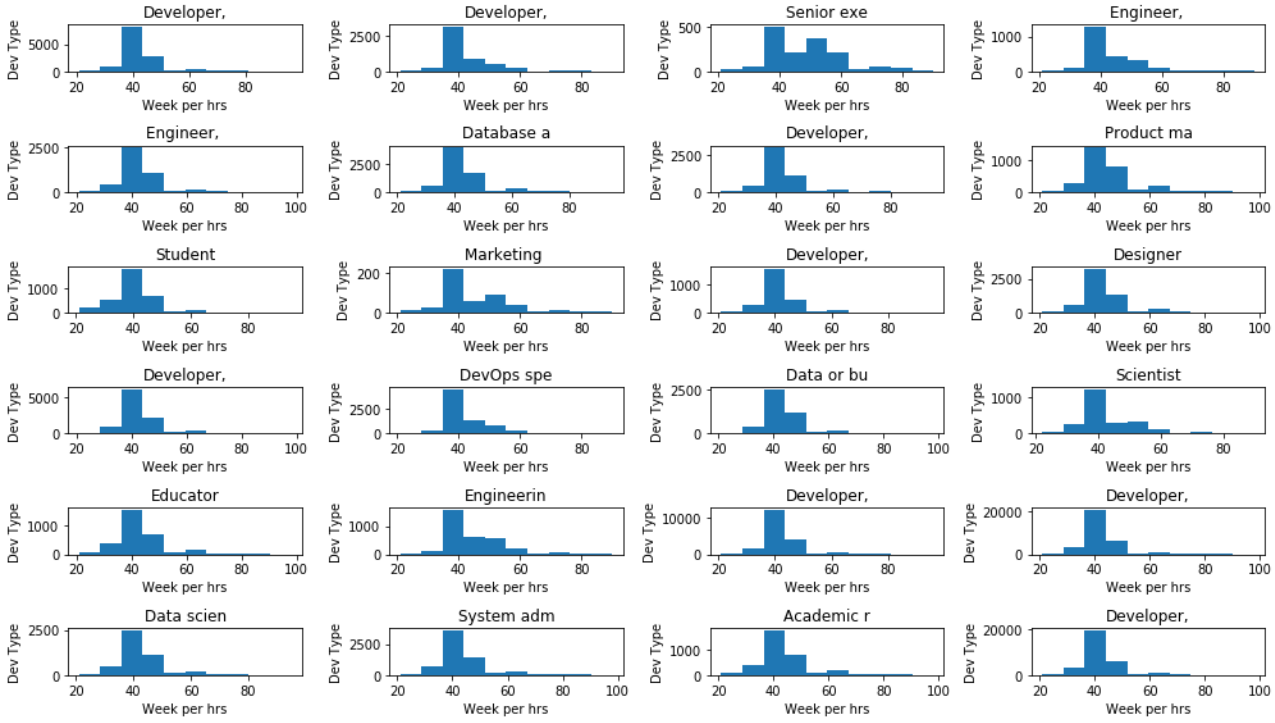


Figura 23.- Histogramas de número de horas de trabajo por tipo de desarrollador.

Por medio de los histogramas se puede observar que los distintos tipos de desarrolladores presentan cantidades de horas de trabajo similares y razonables, en contraste las ingenierías suelen presentar mayores cargas de trabajo quizás influenciado con el hecho de que en dichas empresas suelen tener un horario fijo y mayor que la anteriormente descrita.

Es de destacarse que por medio de su histograma disperso el trabajo con mayor número de horas es el alto ejecutivo (senior executive), esto es muy razonable debido a que este puesto suele ser el encargado de distintas áreas presentes en una empresa y con ello puede estar incluido el hecho de tener que realizar horas extras.

8. Trace histogramas con 10 bins para la edad por género.

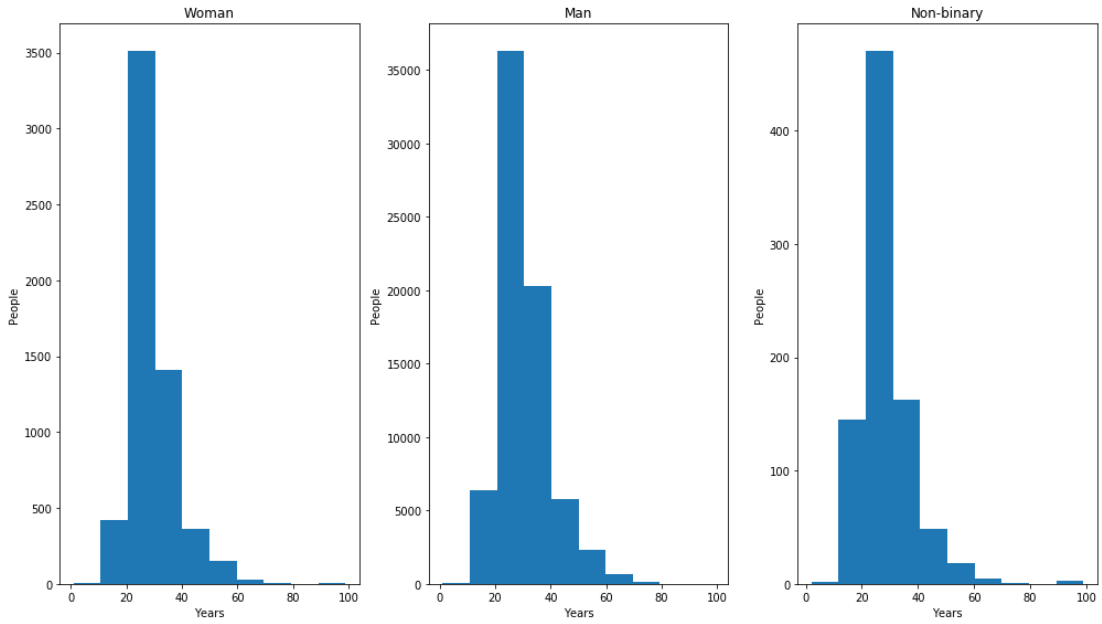


Figura 24.- Histogramas de edad por género.

En primera instancia se puede observar que los histogramas en las tres categorías presenta una forma muy similar, diferenciandose unicament en el número de encuestados en cada uno de ellos, siendo el de los hombres mucho mayor. Por su parte se puede afirmar que el rango de edad más destacado es el comprendido entre los 20 y 40 años, mientras que el mayor número de personas con más de 40 años son hombres.

9. Calcule la mediana, la media y la desviación estándar de la edad por lenguaje de programación

```
HTML/CSS
Median: 28.0
Mean: 29.83861149315315
Standard Deviation: 8.837072940185125
```

```
JavaScript
Median: 28.0
Mean: 30.1186744966443
Standard Deviation: 8.676488395662691
```

```
PHP
Median: 28.0
Mean: 29.400688317983523
Standard Deviation: 8.817271681658973
```

Figura 25 .- Histogramas de número de horas de trabajo por tipo de desarrollador.

Se realizó una primera observación en algunos ejemplos de los lenguajes de tipo web, obteniéndose que las personas que tienen conocimiento y trabajan con ellos están en un rango de 29 a 30 años. Su desviación estándar entre los tres es muy similar y pequeña por lo que la cercanía entre todas estas edades es muy pequeña y con ello se puede afirmar que dicha observación es válida.

```
C
Median: 28.0
Mean: 30.095139607032056
Standard Deviation: 9.118472430968199
```

```
C++
Median: 26.0
Mean: 28.658021405201577
Standard Deviation: 9.837132003254386
```

```
C#
Median: 29.0
Mean: 30.635184057794554
Standard Deviation: 9.204646492094136
```

Figura 26.- Histogramas de número de horas de trabajo por tipo de desarrollador.

Por su parte en los lenguajes de programación de tipo C el rango de edad varía entre los tres. En el lenguaje más viejo, el C, el rango de edad nos dice que personas más mayores son las que aún hacen uso de dicho lenguaje, mientra que C++ el rango de edad es menor en comparación a C++ y C#. Esto nos dice que las personas jovenes están más interesadas en lenguajes más recientes que en los viejos.

```
Python
Median: 28.0
Mean: 29.33033364876828
Standard Deviation: 9.11917739741666
```

```
Java
Median: 28.0
Mean: 29.96342351469866
Standard Deviation: 8.758582417145659
```

Figura 27.- Histogramas de número de horas de trabajo por tipo de desarrollador.

Se realizó una tercera observación pero ahora con diferentes lenguajes muy conocidos, el interés en cuanto a Python y Java son muy similares entre las personas de 28 a 30 años.

10. Calcule la correlación entre años de experiencia y salario anual

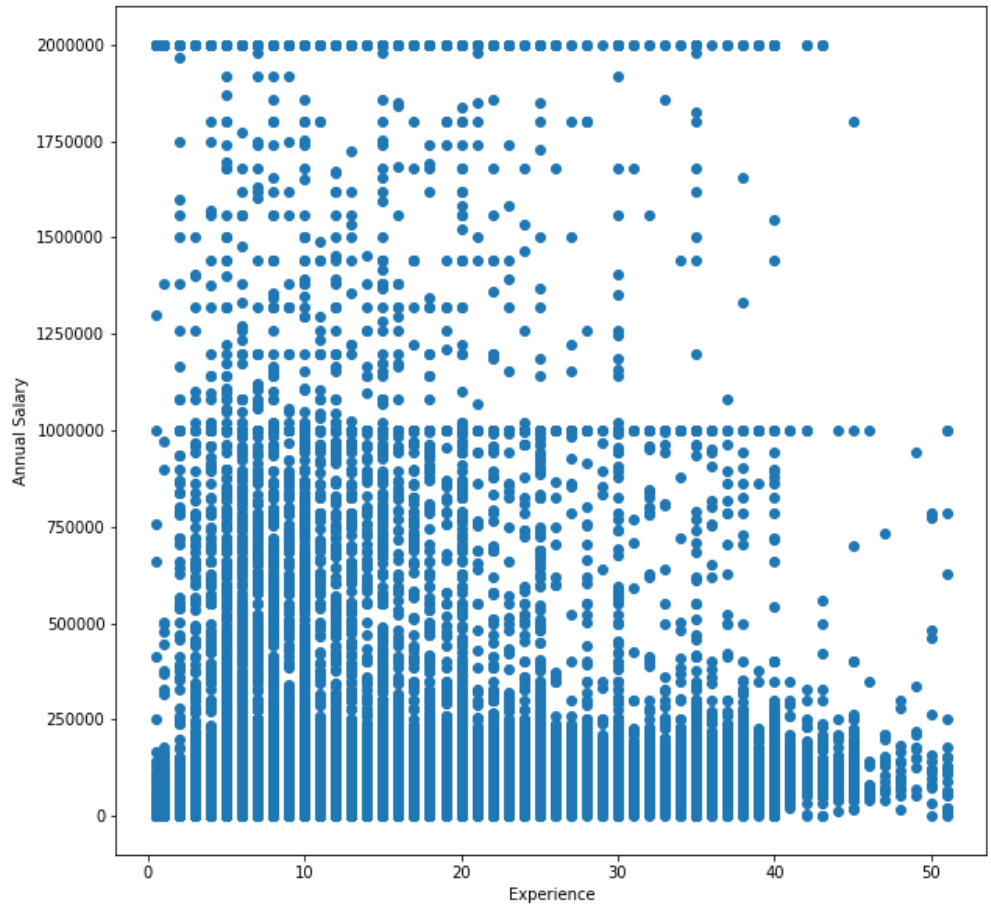


Figura 28.- Histogramas de número de horas de trabajo por tipo de desarrollador.

```
Correlation between years of experience and annual salary: 0.10600459758020611
```

El gráfico nos muestran que la mayoría de los datos se encuentran entre los 5 y 25 años de experiencia, presentando un salario un tanto equitativo entre la franja de los 0 a los 80 mil dolares. No obstante, conforme va aumentando la experiencia el salario se va disminuyendo, esto puede estar influenciado al hecho en que las empresas suelen preferir personas jóvenes debido a sus necesidades como empresa, que podría la utilización de herramientas y lenguajes de programación más recientes. También puede ser que dichas personas ya mayores prefieren ya sea trabajar por su cuenta o jubilarse.

11. Calcule la correlación entre la edad y el salario anual

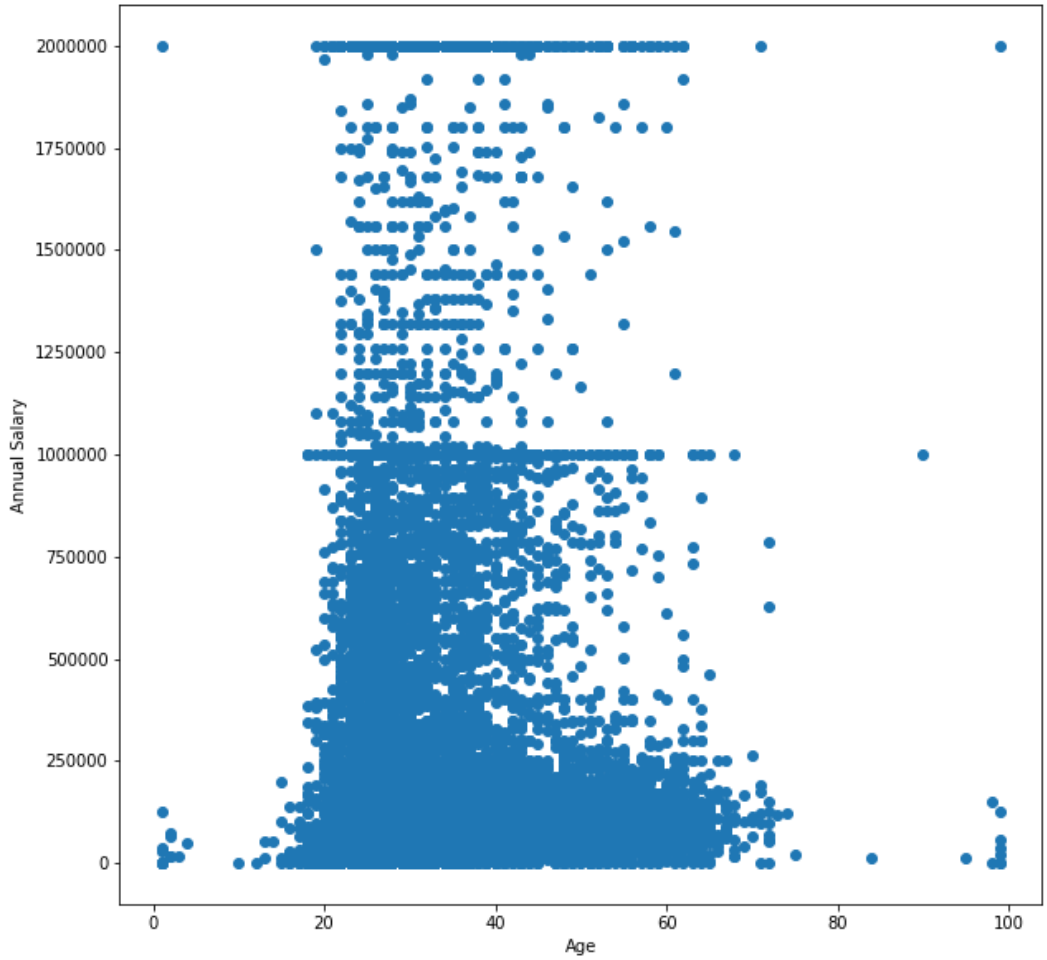


Figura 29.- Histogramas de número de horas de trabajo por tipo de desarrollador.

```
Correlation between age and annual salary: 0.10826846489974247
```

Esta observación confirma lo dicho en el anterior ejercicio, que la mayoría de personas de entre 5 y 25 años de experiencia son jóvenes, los cuáles dominan la cantidad de respuestas en la encuesta. La regularidad en cuanto a salario se mantiene entre la edad de 20 a 40 años, la cual va desde los 0 a los 80 mil dolares aproximadamente. Dicha regularidad va dismuyendo a partir de los 45 años, que es cuando tanto la cantidad de respuestas como el salario anual se va disminuyendo.

12. Calcule la correlación entre el nivel educativo y el salario anual. En este caso, reemplace la cadena del nivel educativo por un índice ordinal (por ejemplo, Primaria = 1, Secundaria = 2, etc.).

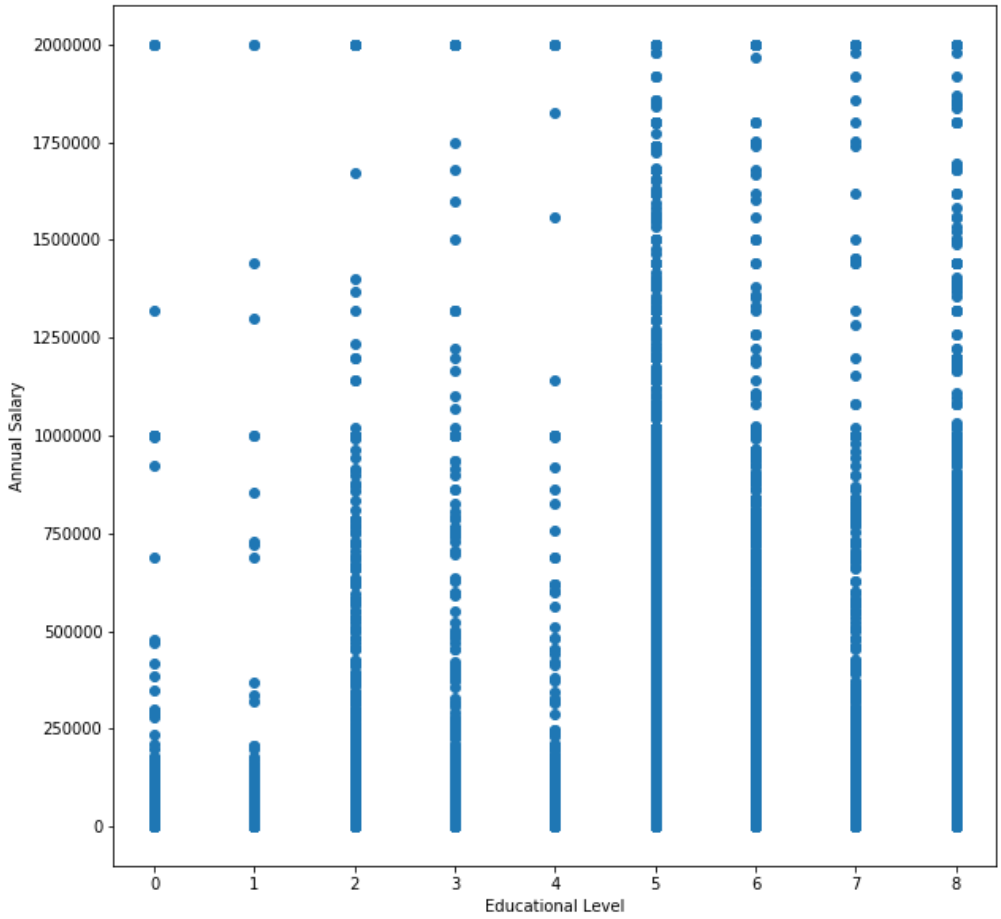


Figura 30.- Histogramas de número de horas de trabajo por tipo de desarrollador.

```
Correlation between years of expenrience and annual salary:0.004743429243741032
```

El grafico muestra salarios anuales que en su mayoría mantienen regularidad entre un dato y el otro. Se puede observar que entre mayor sea el nivel escolar mayor es la renumeración, aunque en el nivel 4, el cual corresponde al Professional Degree, el salario anual ve seriamente disperso e irregular.

Se debe destacar el hecho de que el nivel 5, el cual es es el grado de bachillerato, se ve una equidad salarial muy regular, manteniéndose muy poco dispersos los datos del uno al otro. Por su parte, en el nivel 6, el cual describe un grado universitario o colegial, los datos a partir de los 100 mil dolares se van dispersando, describiendo así que pocos son los que mantienen un salario más alto que esa cantidad. Tomando en consideración ambos niveles, dichos datos quieren decir que no es necesario un grado universitario para tener un buen salario y que la mayoría de los conocimientos necesarios para ser desarrollador fueron aprendidos en el grado de bachillerato.

13. Obtenga un diagrama de barras con las frecuencias de los diferentes lenguajes de programación.

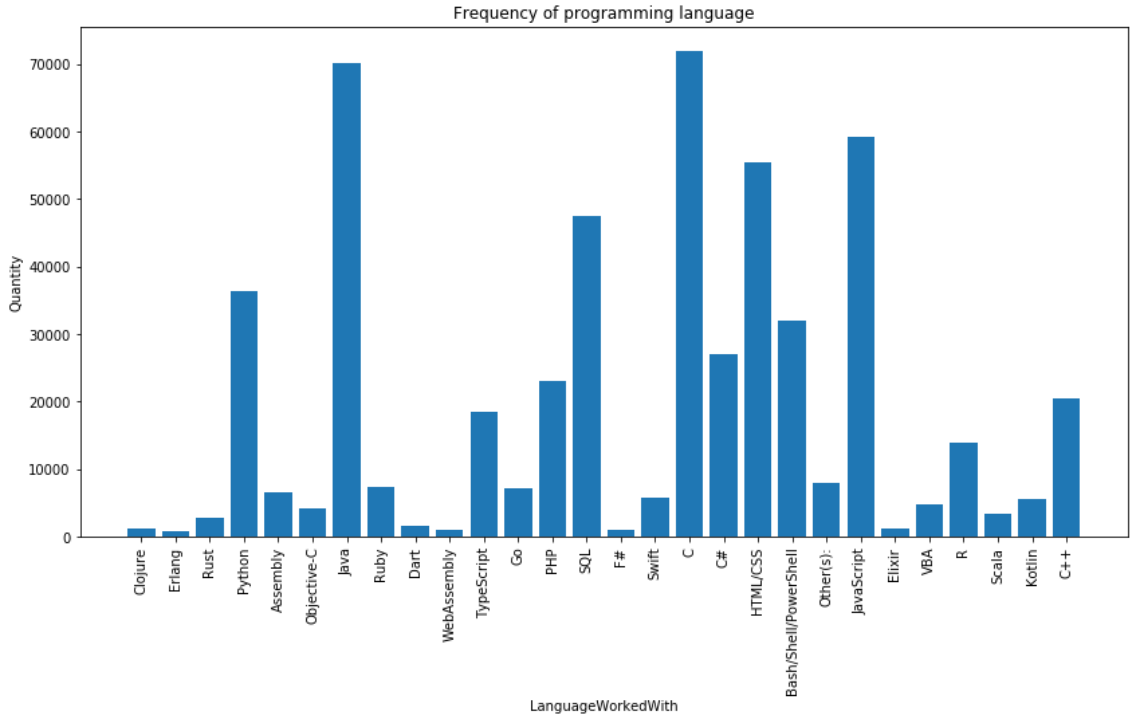


Figura 31.- Histogramas de número de horas de trabajo por tipo de desarrollador.

El gráfico muestra que los lenguajes más populares entre los desarolladores son dos que a pesar de tener ya tiempo todavían son utilizados, los cuales son C y Java. Subsiguientemente están lenguajes que son utilizados en el desarrollo web: HTML/CSS, Javascript y SQL. Después están los que a pesar de no aparecer en un lugar más alto, van tomando ventaja respecto al tiempo: Python, C# y PHP. Por último, y como debía esperarse, están lenguajes más antiguos y complejos como lo son Assembly y F#.

Conclusiones:

Después de las observaciones que se hicieron en cada uno de los ejercicios ha

- Aún existe una inequidad de género dentro del área de la programación y el desarrollo de software, en donde dicha industria aún es dominada por los hombres en comparación a los números de mujeres y otras minorías presentes.
- En cuanto al salario anual y el género, no existe una gran brecha salarial entre las tres categorías.
- Si existe una brecha salarial muy evidente respecto a la etnia y al país en donde se reside, siendo que en países considerados desarrollados la renumeración es mayor en comparación a los países en vías de desarrollo.
- La edad de una persona de igual forma puede afectar en cuanto al salario que recibe, siendo mayormente afectados las personas mayores a 40 años, en donde su salario se ve reducido. Inversamente suele verse el resultado contrario en cuanto a la experiencia en programación, siendo el rango de 5 a 25 años de experiencia los que presentan un salario más estable y regular.
- El horario de trabajo y el salario anual son muy similares entre las áreas que tienen algo en común, como lo es la ingeniería, el área de la educación o el desarrollo web pero en cuanto a áreas diferentes se ve un cambio considerable, esto debido a que los horarios de trabajo suelen ajustarse a las necesidades de la empresa. En algunos trabajos no se requiere tanto tiempo extra mientras que en otros es algo necesario.