

Galileo
UNIVERSIDAD
La Revolución en la Educación

Departamento de Ingeniería

Universidad Galileo

Postgrado en Sistemas de Información

Proyecto

Modelos y simulación

Integrantes:

Santos López Tzoy - 15002241

Hery Fernando - 13002242

Catedrático:

Dr. Erick Reyes

Guatemala, 02 de abril de 2023

<i>Introducción</i>	4
<i>Objetivos del proyecto.....</i>	5
<i>Base de datos utilizado.....</i>	5
<i>Repositorio de todas las bases de datos usado en el proyecto, powerpoint, etc.</i>	5
Cantidad de datos utilizado fue suficiente	6
<i>Requerimientos para analizar la base de datos.....</i>	6
Colab Research	6
Instalación de Docker en MacOS.....	7
Instalación de librerías en Docker en MacOS	7
Correr R y Colab de forma local en MacOS	8
Cargar archivo de R en Jupyter	10
<i>Modelos utilizados en el proyecto.....</i>	10
<i>Primer modelo construido de forma equivocada.....</i>	10
¿Qué significa que el resultado de una correlación es negativo?.....	11
Primera matriz de correlación con errores	12
<i>Segundo modelo de forma incorrecta:</i>	13
Correlación entre variables.....	14
¿Qué inconvenientes había con el segundo análisis?	14
Variables de entrada utilizado:	14
<i>Tercer modelo construido de forma correcta</i>	15
Resultado de matriz de correlación.....	15
Datos estadísticos.....	15
Distribución de frecuencias	17
Distribución normal.....	19
Distribución de Poisson	20
<i>Cuarto modelo construido de forma correcta por Hery</i>	21
Peones agropecuarios	21
Vendedores de productos para higiene personal	23
Personal de Limpieza	25
Obreros de la construcción	27
Profesionales de la enseñanza	28
Peones de la minería	30
<i>Trabajo realizado por cada estudiante</i>	32

Estudiante: Hery Fernando Gomez de Leon	32
Estudiante: Santos López Tzoy	32
<i>Descargar software de Simulación @RISK.....</i>	33
<i>Conclusiones.....</i>	34
<i>Referencias bibliográficas.....</i>	34

Introducción

En el presente trabajo se analiza ¿Cómo el nivel académico puede influir para que una persona tenga un mayor salario? Para el análisis se utilizó la base de datos de <https://www.ine.gob.gt/encuesta-nacional-de-empleo-e-ingresos/> y en esta base de datos se construyó uno nuevo con datos de las variables del interés del proyecto.

La variable de salida “SUELDO O SALARIO” y las variables de entrada “EDAD”, “TAMAÑO EMPRESA”, “ESCOLARIDAD”, “GÉNERO”, etc. Por ejemplo: en base a la edad y el nivel académico puede influir para que un trabajador obtenga un mejor salario.

En el documento están todos los modelos que se construyeron al inicio del proyecto y que posteriormente fueron descartados en base a muchos inconvenientes como: mal análisis del modelo, objetivos del proyecto no establecidos, variables no consideradas para el estudio, etc.

Objetivos del proyecto

- Averiguar si en base al grado académico esto influye para que una persona obtenga un mejor
- Conocer si los trabajadores en Guatemala, en su salario son competitivos con otros países de Centro América.

Base de datos utilizado

Para la elaboración del proyecto, se utilizó la base de datos disponible en <https://www.ine.gob.gt/encuesta-nacional-de-empleo-e-ingresos/>. Las bases de datos que se muestran están contenidas las base de datos de forma separada, por ejemplo: el del año 2019 en una archivo distinto al del 2021, sin embargo para el proyecto fue necesario hacer la unión de ambas bases de datos y está puede ser encontrada en el repositorio de GitHub <https://github.com/santoslopez/modelosysimulacion>. En este repositorio aparecen ambas bases de datos unidos.

En el sitio de INE el formato en que está disponible en la base de datos es en Excel y CSV. Ambos formatos de archivo pueden ser cargados a algún lenguaje de programación como Python, Julio, R, etc., para analizar.

The screenshot shows the official website of the Instituto Nacional de Estadística de Guatemala (INE). The header includes the INE logo, contact information (comunicacion@ine.gob.gt, +502 2315 4700), and social media links. The main navigation menu has options like Inicio, Acerca del INE, Estadísticas por tema, Servicios estadísticos, SINACIG, Información Pública, Bolsa de Empleo, Contáctenos, and Solicitudes. Below the menu, a breadcrumb trail shows the path: Inicio > Estadísticas por tema > Empleo. On the left, there's a sidebar with links for Publicaciones, Indicadores, Bases de datos (which is highlighted), and ENEI 2021. The main content area is titled 'Encuestas de hogares y personas' and 'Encuesta Nacional de Empleo e Ingresos'. It contains a descriptive text about the survey's objectives and coverage, followed by a form to select the year and month. At the bottom, there's a Mac OS X-style dock with various application icons.

Ilustración 1. Base de datos INE

Repositorio de todas las bases de datos usado en el proyecto, powerpoint, etc.

La base de datos utilizado puede ser visto en el siguiente repositorio de GitHub <https://github.com/santoslopez/modelosysimulacion> en la carpeta llamado BD, en este

repositorio está además el código utilizado en R para analizar los datos en excel, la presentación en power point, etc.

Cantidad de datos utilizado fue suficiente

Uno de los mayores problemas con los que uno se enfrenta es son los datos, pueden faltar datos, tener muchos datos y estar en casi en la cantidad adecuada de datos.

En la siguiente imagen del Excel aparece la base de datos utilizado y está corresponde a los años 2019 y 2021.

The screenshot shows a Microsoft Excel spreadsheet titled "2001_2018". The data starts at row 2 and ends at row 36. The columns are labeled from A to Z. The data consists of various numerical values and some text entries. The Excel ribbon is visible at the top, and the status bar at the bottom indicates "Último: Accesibilidad es necesario investigar".

Ilustración 2. Base de datos utilizado con los años 2021 y 2019

En esta base de datos, están contenidos los años 2019 y 2021 juntos, y la cantidad de filas de la base de datos son 48639

This screenshot shows the same Excel spreadsheet as the previous one, but it appears to be a different view or a later version. The data structure is identical, with rows 2 to 36 and columns A to Z. The Excel ribbon and status bar are visible.

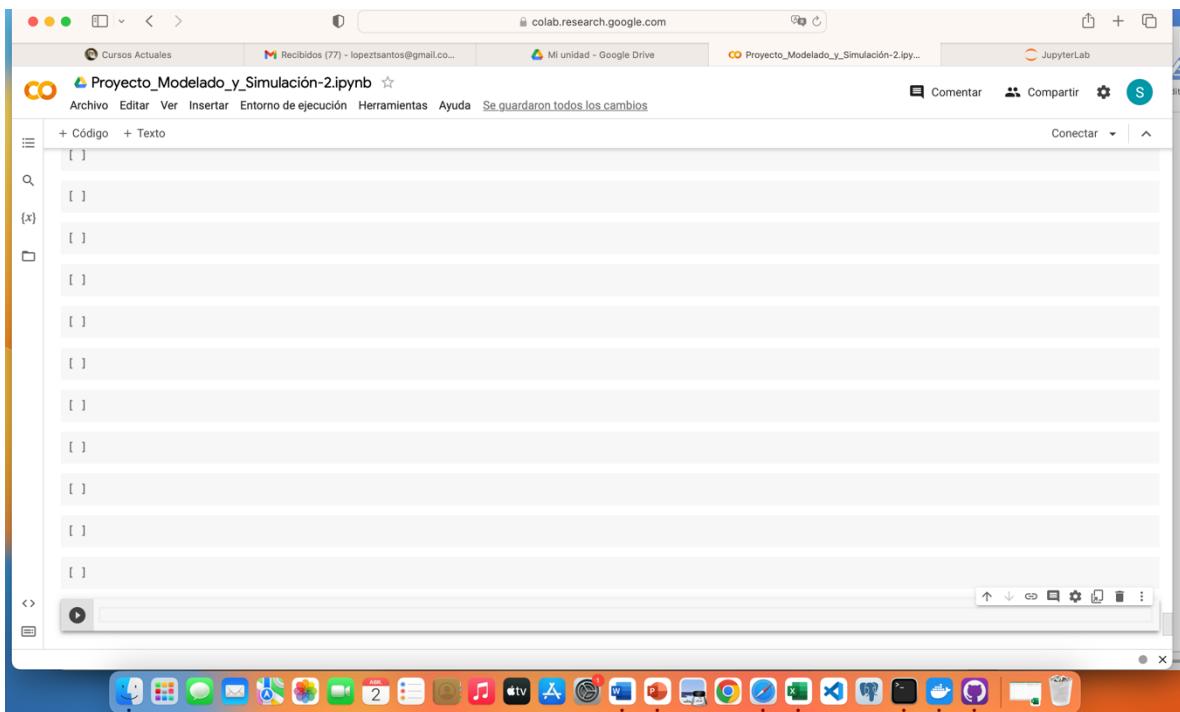
Requerimientos para analizar la base de datos

Los archivos de Excel pueden ser analizados utilizando Python, Julia o R. Para el proyecto, se utilizó R y Colab. R y Colab pueden ser instalados en Windows, Ubuntu o MacOS.

Para correr R y Colab en MacOS, es necesario instalar Docker y luego ejecutar unos comandos en MacOS para que el contenedor descargue unas librerías necesarias.

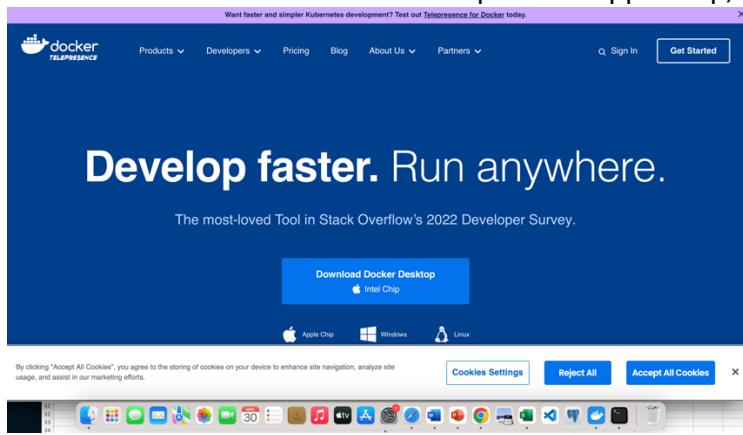
Colab Research

Para analizar la base de datos puede ser usado <https://colab.research.google.com>. Básicamente es un servicio en línea en donde uno puede utilizar R, Python, etc., para trabajar de forma gratuita.



Instalación de Docker en MacOS

Es necesario ir a la página <https://www.docker.com> elegir el sistema operativo correspondiente. Si cuenta con una MacBook con chip Intel o Apple chip, Linux o Windows.



El siguiente comando es para instalar Jupiter sin necesidad de instalar MacOS. Para más información visitar la siguiente página <https://jupyter.org/install>

Instalación de librerías en Docker en MacOS

En la siguiente imagen aparece el comando “docker pull jupyter/datascience-notebook” esté se ingresa a la terminal para que termine de descargar la imagen del docker hub. En este comando le indicamos que la instalación que nos interesa es la de jupyter y con unas librerías.

```

Last login: Tue Mar 21 18:16:52 on console
santoslopeztzoy@192 ~ % docker run -p 8888:8888 jupyter/dat...
santoslopeztzoy@192 ~ % jupyter notebook

zsh: command not found: jupyter
santoslopeztzoy@192 ~ % docker pull jupyter/datascience-notebook
zsh: command not found: docker
santoslopeztzoy@192 ~ % docker pull jupyter/datascience-notebook
Using default tag: latest
latest: Pulling from jupyter/datascience-notebook
b2ddfd3d37773: Pulling fs layer
9343f5e9ff5: Pulling fs layer
47f6c476a899: Pulling fs layer
5866cccf32e4: Pulling fs layer
d547e74cb492: Pull complete
204c6c819856: Pull complete
64f6d643829b: Pull complete
474fb700ef54: Pull complete
e85cabbd117e: Pull complete
6e3f89464dc0: Pull complete
14123a6d5249: Pull complete
30a45828d00c: Pull complete
ef1c81526c92: Pull complete
6dc28eebbe2e: Pull complete
222036d8dc53: Pull complete

```

Ilustración 3. Instalar librerías en Docker

Correr R y Colab de forma local en MacOS

Para levantar el docker y poder utilizar R y colab de forma local es necesario ingresar el siguiente comando en la terminal: docker run -p 8888:8888 jupyter/datascience-notebook

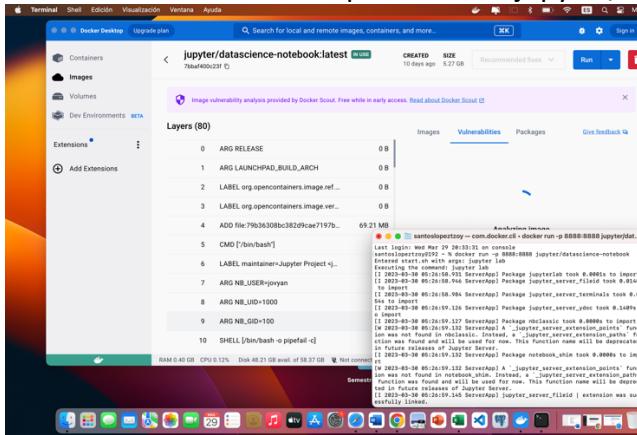


Ilustración 4. Ejecutando Jupyter

Cada vez que uno levanta el docker, es necesario copiar el código que aparece en el texto token. Por ejemplo: en la imagen aparece un código similar a 9797b6c13d9695b35377c408603568c0edb475dd. Este número es generado cada vez que se inicia docker y debe ser copiado en la siguiente interfaz:

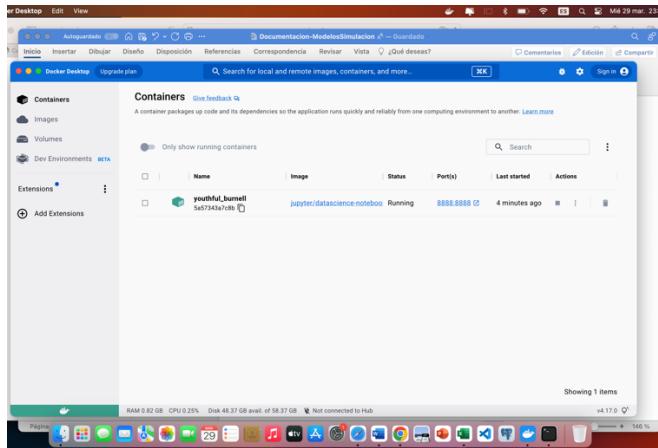
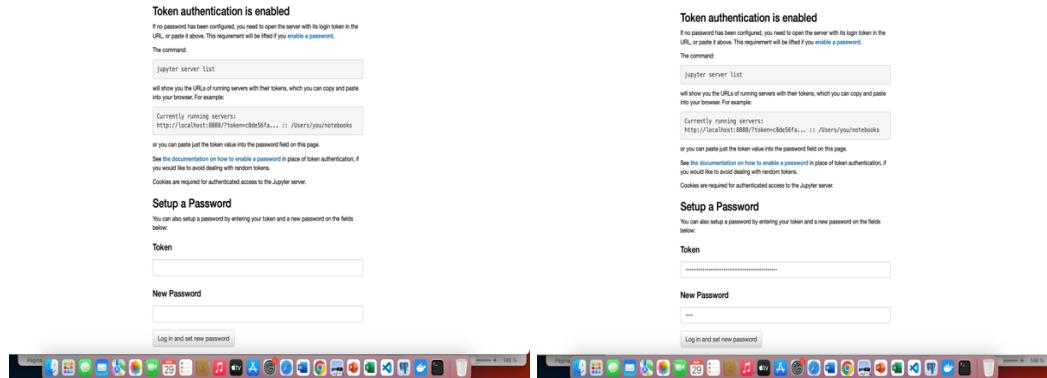


Ilustración 5. Abrir Jupyter de forma local

Al darle al enlace <http://localhost:8888/login?next=%2Flab%3F> se abre la siguiente ventana:



En la imagen anterior, en el input token hay que copiar el código generado al ejecutar el comando: “`docker run -p 8888:8888 jupyter/datascience-notebook`” y colocar cualquier clave en el Input New Password. El token que se colocó aparece en la terminal y se colocó una contraseña. Ahora hay que darle clic al botón Log in and set new password. Ahora se genera una interfaz en donde uno puede seleccionar si desea utilizar Python 3, Julia 1.8.5, R, etc. En este caso seleccionamos R.

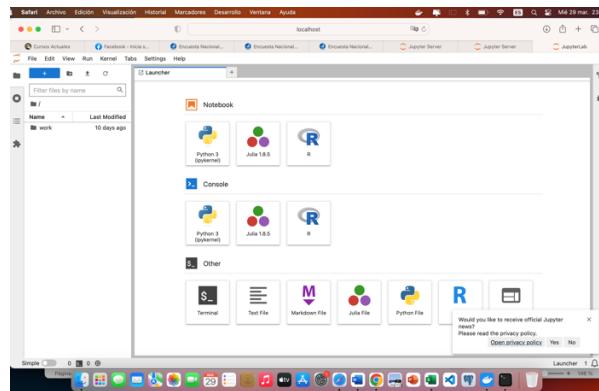


Ilustración 6. Elegir lenguaje de programación.

Cargar archivo de R en Jupyter

Al tener un archivo en R creado, es posible subirlo al entorno donde se está trabajando.

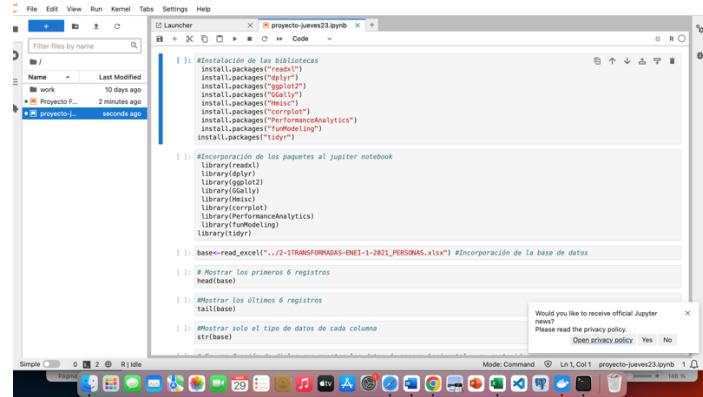


Ilustración 7. Cargar archivo de R en Jupyter

Modelos utilizados en el proyecto

En el proyecto se consideraron 3 modelos.

El primer modelo la base de datos los errores que tenía son:

- No estaba limpia, se consideraron valores vacíos y esto afectó en los resultados en la correlación, en la primera matriz de correlación obtenida se encontró con muchos valores NA
 - Muchas variables de entrada no eran relevantes, sin embargo, en el primer modelo estaban siendo considerados.
 - Se consideraron todos los empleos, no había un filtro para indicar solo cierto empleo.

En el segundo modelo se encontró con los siguientes inconvenientes:

- Se escogió un solo empleo, por ejemplo: Ingeniero, sin embargo el inconveniente está en que toma en cuenta todos los tipos de Ingenieros que existen como Ingeniero en Mecatrónica, Ingeniero eléctrico, Ingeniero Civil, etc., esto provoco que algunos valores de la matriz de correlación no fueran los esperados.

En el tercer modelo utilizado se realizó:

- Correcciones encontradas en los primeros 2 modelos.
 - Solo se tomó en cuenta un tipo de profesión, por ejemplo: Ingeniero en Sistemas de Computación del Gobierno de Guatemala.
 - No se tomaron en cuenta valores que tengan NA, para esto fue necesario hacer una limpieza en R de la base de datos.

Para más detalles en las siguientes secciones, se profundiza un poco más sobre los modelos utilizados.

Primer modelo construido de forma equivocada

En el primer modelo construido, la variable de salida es “SALARIO DEL TRABAJADOR”, está es la variable de interés y se correlacionó con otra variable. Por ejemplo: “SALARIO DEL

TRABAJADOR” y como segunda variable (dominio de estudio, tamaño de empresa, alimentación, edad, tiempo de trabajo, años, meses, jornada, etc.)

[¿Qué significa que el resultado de una correlación es negativo?](#)

La relación entre la variable de interés: sueldo o salario y la segunda variable la relación entre ellas es inversamente proporcional, esto quiere decir que si en una variable hay un aumento en la otra existe una disminución.

En la siguiente tabla se muestran los resultados de la correlación del primer modelo utilizado. En este primer modelo se consideraron variables equivocadas como hora lunes, hora martes, hora miércoles, hora jueves, hora viernes, horas sábado, horas domingo.

Variable de 2da variable interés: SUELDO DEL TRABAJADOR	Resultado correlación
Dominio de estudio	-0.3104456
Tamaño de empresa	0.1360462
Alimentación	0.125341
Edad	0.2151899
Tiempo de trabajo	0.08418937
Años	0.2240756
Meses	0.01114745
Jornada	0.04579173
Sexo	-0.009851248
Hora lunes	0.09956595
Horas martes	0.09765284
Hora miércoles	0.08990469
Horas jueves	0.102946
Horas viernes	0.1241907
Horas sábado	-0.1657717
Horas domingo	-0.029703
Segundo trabajo	0.06027297
Cambio de trabajo	0.2059654

¿Cuál es el problema con esté análisis inicial?

- Los resultados de la correlación mostrada en la tabla toman en cuenta todos los empleos como: guardias de seguridad, secretarias, Ingenieros, Arquitectos, etc., por lo tanto, fue necesario repetir el modelo porque el análisis no es el correcto. Para nuestro modelo solo queremos considerar un solo empleo. Por ejemplo: conocer en base a sus estudios que tanto puede ganar un Ingeniero en Sistemas de Computación.
- En la base de datos se consideraron valores vacíos y esto influía demasiado en los resultados de la correlación.

[Primera matriz de correlación con errores](#)

La primera matriz de correlación utilizada es la que aparece en la siguiente tabla. En esta matriz de correlación, no se utilizó las variables adecuadas para correlacionar debido que aparecen muchos “NA” y algunas correlaciones están en negativo. Por lo tanto, para crear el modelo del proyecto, fue necesario volver a analizar los datos que aparecen en Excell.

	SUELDO O SALARIO	DOMINIO DE ESTUDIO	SEXO	EDAD	LUGAR DE TRABAJO	ALFABETISMO
SUELDO O SALARIO	1	NA	NA	NA	NA	NA
DOMINIO DE ESTUDIO	NA	1	-0.008083521	-0.1022758	NA	NA
SEXO	NA	-0.008083521	1	0.04117599	NA	NA
EDAD	NA	-0.1022758	0.0411758	1	NA	NA
LUGAR DE TRABAJO	NA	NA	NA	NA	1	NA
ALFABETISMO	NA	NA	NA	NA	NA	1

	SUELDO O SALARIO	DOMINIO DE ESTUDIO	SEXO	EDAD	LUGAR DE TRABAJO	ALFABETISMO
SUELDO O SALARIO	1	NA	NA	NA	NA	NA
DOMINIO DE ESTUDIO	NA	1	-0.008083521	-0.1022758	NA	NA
SEXO	NA	-0.008083521	1	0.04117599	NA	NA
EDAD	NA	-0.1022758	0.0411758	1	NA	NA
LUGAR DE TRABAJO	NA	NA	NA	NA	1	NA
ALFABETISMO	NA	NA	NA	NA	NA	1

Segundo modelo de forma incorrecta:

Para la elaboración del segundo modelo fue necesario realizar:

- Limpiar la base de datos, verificar que variables de entrada están vacías o tienen NA. Las variables que están vacías no tomarlas en cuenta, para esto previamente se realizó una limpieza utilizando R.

```

File Edit View Run Kernel Tabs Settings Help
+ X Code
Filter files by name
Name Last Modified
work 11 days ago
2-1TRANS... 9 hours ago
2-1TRANS... 9 hours ago
Proyecto_... 10 hours ago
Proyecto_... 2 minutes ago

[53]: base2 <- read_excel("2-1TRANSFORMADAS-ENEI-1-2021_PERSONAS.xlsx") #Incorporación de la base de datos
columnas_na <- c("SUELDO O SALARIO","SEXO","EDAD") # reemplaza esto por los nombres de las columnas que deseas evaluar
datos_sin_na <- base2[complete.cases(base2[columnas_na]), ]
head(datos_sin_na)

A tibble: 6 x 261
DOMINIO DE ESTUDIO AREA UPM FACTOR DE EXPANSIÓN CSUM(HOG) CÓDIGO DE LA PERSONA EN LA BOLETA SEXO EDAD PARENTESCO PUEBLOS ... USO DE RECURSOS ALIMENTOS CANTI QUETZ/
<dbl> <dbl>
1 1 26 295 1 1 2 43 1 3 ... NA 2
1 1 26 295 1 2 1 20 3 3 ... NA NA
1 1 26 295 2 4 2 24 3 3 ... NA NA
1 1 26 295 3 2 1 33 13 3 ... NA NA
1 1 26 295 4 1 1 66 1 3 ... NA 2
1 1 26 295 6 1 1 56 1 3 ... NA 2

```

[54]: correlation2 <- cor(datos_sin_na\$"SUELDO O SALARIO", datos_sin_na\$"EDAD")
print(correlation2)

[1] 0.2151899

Ilustración 8. Código del proyecto por Hery

Correlación entre variables

Para la correlación entre la variable de interés “SUELDO O SALARIO” con otras variables de salida, la correlación es positiva.

Sueldo o salario, variable de interés	Variable de entrada	Resultado correlación
	EDAD	0.9153992
	AÑOS	0.5834472
	DOMINIO DE ESTUDIO	0.4491378

¿Qué inconvenientes había con el segundo análisis?

En la tabla anterior pareciera que todo iba bien, sin embargo, nuevamente al elegir un tipo de trabajo no se clasifico de la forma correcta. Ejemplo: se consideró el trabajo de Ingenieros, sin embargo, hay muchas subdivisiones: Ingeniero de mecatrónica, Ingeniero Aeroespacial, Ingeniero en Sistemas de Computación, para resolver el problema, es necesario seleccionar solo una subdivisión de trabajo.

Variables de entrada utilizado:

Se analizo una gran variedad de variables de entrada, por ejemplo: dominio de estudio, sexo, edad, lugar de trabajo, alfabetismo, etc.

Tercer modelo construido de forma correcta

En el tercer modelo se realizó lo siguiente:

- Se corrigieron las observaciones de la retroalimentación obtenidas por el catedrático del curso de Modelos y Simulación detectadas en los primeros 2 modelos utilizados.
- Se utilizó una base de datos en donde estaba contenido los datos del año 2019 y 2021.
- Se limpió la base de datos para utilizar solo los datos que no tengan NA.

Resultado de matriz de correlación

En la siguiente tabla se muestran los resultados de la matriz de correlación, la variable de interés es salario del trabajador. En muchas pruebas realizadas se consideraron distintos empleos, y dependiendo del tipo de trabajo, el valor de la matriz de la correlación cambió bastante, por ejemplo, hay empleos como: Secretaría Bilingüe en donde las mujeres ganan mejor que el hombre. El nivel académico también es determinante para que un trabajador obtenga una mejor remuneración.

	Salario del trabajador	Edad	Años en la empresa	Tamaño de la empresa	Nivel académico	Sexo
Salario del trabajador	1	0.09769185	0.14581873	0.07917097	0.27941675	-0.03228390
Edad	0.09769185	1	0.28592124	0.06710769	-0.05245466	0.02303865
Años en la empresa	0.14581873	0.28592124	1	0.01473353	-0.24088899	-0.03546251
Tamaño de la empresa	0.07917097	0.06710769	0.01473353	1	0.02360268	0.02716183
Nivel académico	0.27941675	-0.05245466	-0.24088899	0.02360268	1	-0.02234456
Sexo	-0.03228390	0.02303865	-0.03546251	0.02716183	-0.02234456	1

Datos estadísticos

En la tabla sueldo del trabajador se concluye que los trabajadores, la mayoría gana Q.3000.00.

SUELDO DEL TRABAJADOR	
Media	2426.33709981168
Mediana	2600
Desviación estándar	1227.05581108106
Moda	3000

Varianza	1505665.9635078
Primer cuartil	1500
Segundo cuartil	2600
Tercer cuartil	3000
Percentil 25	1500
Percentil 50	2600
Percentil 75	3000

Tabla 1. Salario del trabajador.

En la siguiente tabla se puede observar que el género que más se repite son los hombres.

SEXO	
Media	1.54785020804438
Mediana	2
Desviación estándar	0.498050599591476
Moda	2
Varianza	0.248054399753429
Primer cuartil	1
Segundo cuartil	2
Tercer cuartil	2
Percentil 25	1
Percentil 50	2
Percentil 75	2

Tabla 2. Género del trabajador

En la tabla edad, la mayoría de los jóvenes tienen 18 años.

EDAD	
Media	33.7184466019417
Mediana	32
Desviación estándar	18.4104380364703
Moda	18
Varianza	338.944228694714
Primer cuartil	20
Segundo cuartil	32
Tercer cuartil	45
Percentil 25	20
Percentil 50	32
Percentil 75	45

Tabla 3. Tabla de edad

NIVEL DE ESTUDIOS APROBADOS	
Media	2.98474341192788
Mediana	3
Desviación estándar	1.11357293924357
Moda	4

Varianza	1.24004469101556
Primer cuartil	2
Segundo cuartil	3
Tercer cuartil	4
Percentil 25	2
Percentil 50	3
Percentil 75	4

Tabla 4. Nivel de estudios aprobados.

AÑOS LABORANDO	
Media	7.25936199722608
Mediana	5
Desviación estándar	8.12801220287992
Moda	0
Varianza	66.0645823701649
Primer cuartil	2
Segundo cuartil	5
Tercer cuartil	10
Percentil 25	2
Percentil 50	5
Percentil 75	10

Tabla 5. Años laborando.

Distribución de frecuencias

Las siguientes distribuciones de frecuencia se utilizan para conocer la cantidad de veces que se repite la edad, años de trabajo, sueldo del trabajador, etc. Por ejemplo, en la distribución por edad, aparece que los que tienen 25 años hay 150 personas trabajando con esta edad, las personas jóvenes son los que más trabajan.

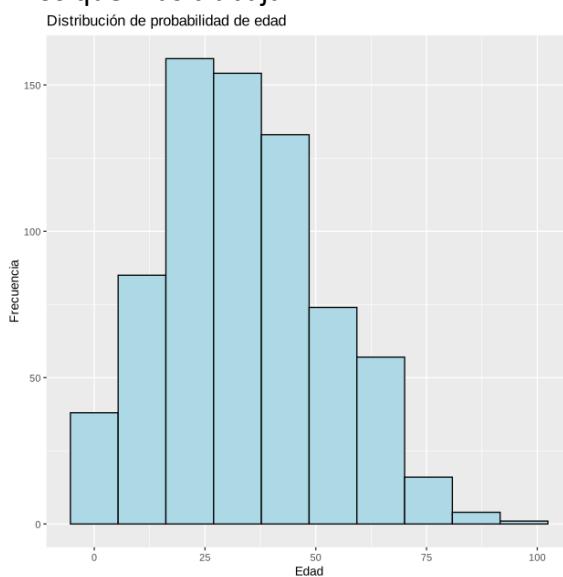


Ilustración 9. Distribución de frecuencias por edad

Son pocas las personas que trabajan muchos años en la empresa. La frecuencia que más se repiten son años de trabajo de 0 a 12 años de trabajo.

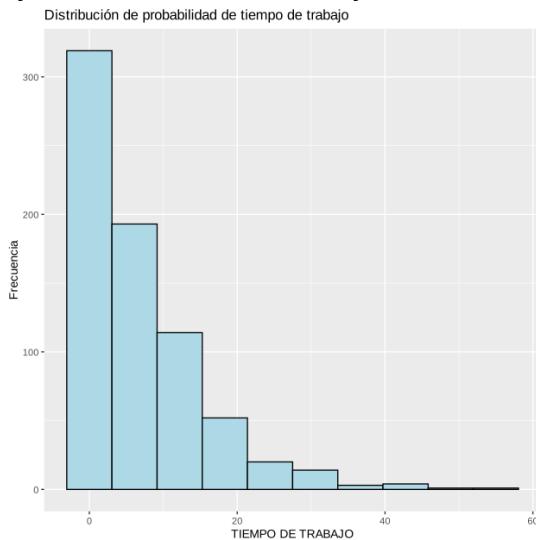


Ilustración 10. Distribución de frecuencias por tiempo de trabajo.

El tamaño de la empresa la frecuencia más alta está alrededor de 700 empleados laborando

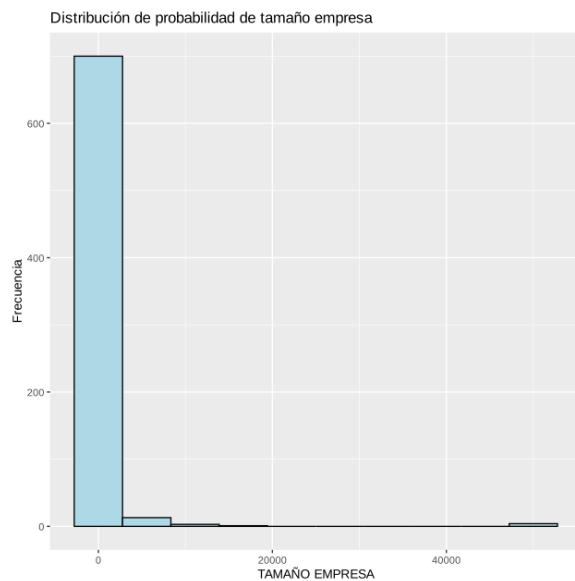


Ilustración 11. Distribución de frecuencias por tamaño de empresa.

En la siguiente imagen, para este tipo de trabajo, son pocos los que ganan Q.5,000.00, el salario del trabajador está entre los Q.2,500.00 a Q.3,500.00

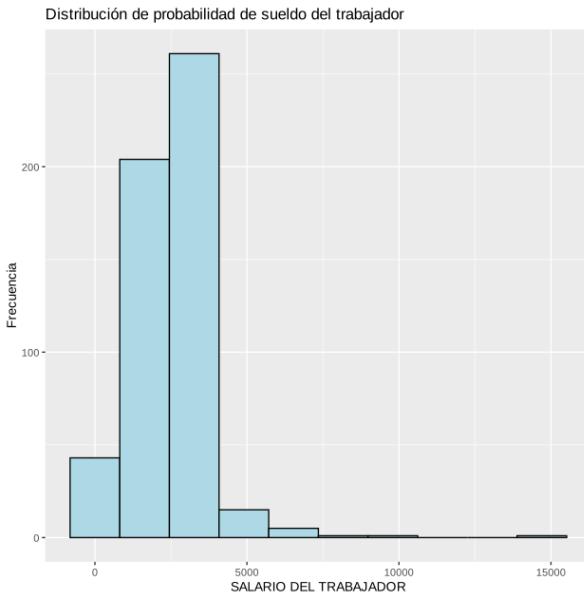


Ilustración 12. Distribución de frecuencias por salario del trabajador.

En la siguiente ilustración, aparece que muchas personas su nivel académico es demasiado bajo y otras personas están en un académico a nivel diversificado o universitario.

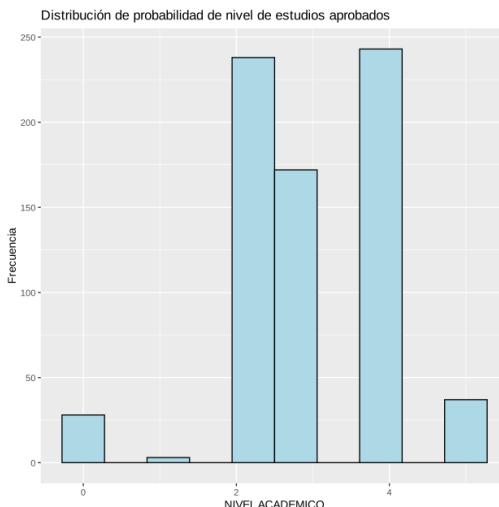


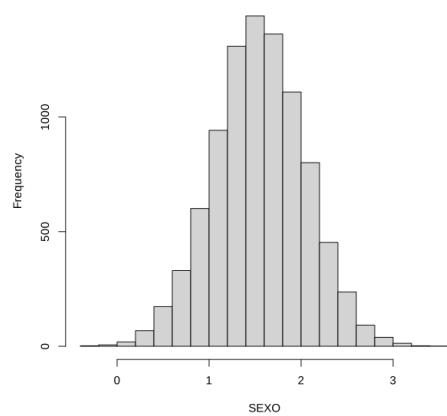
Ilustración 13. Distribución de frecuencias de nivel académico.

Distribución normal

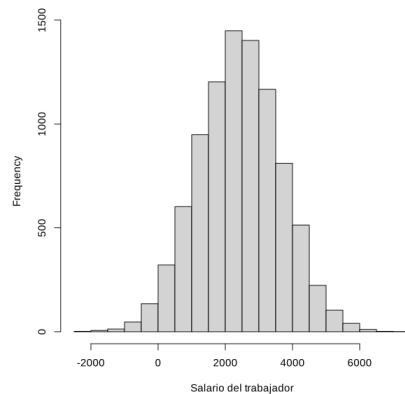
En las siguientes imágenes están la distribución normal para las variables: sueldo del trabajador, sexo, años trabajando en la empresa, etc.

Para sacar la distribución normal, es necesario conocer previamente los valores de la desviación estándar y la media. Estos valores son utilizados en la distribución normal para generar X cantidad de valores aleatorios que uno indique para obtener la distribución de frecuencias. En el proyecto el valor de X=9000

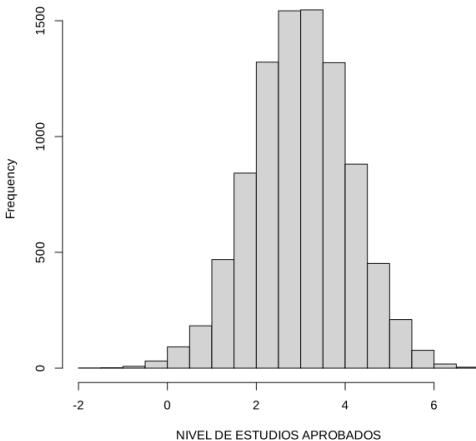
Distribución normal de SEXO



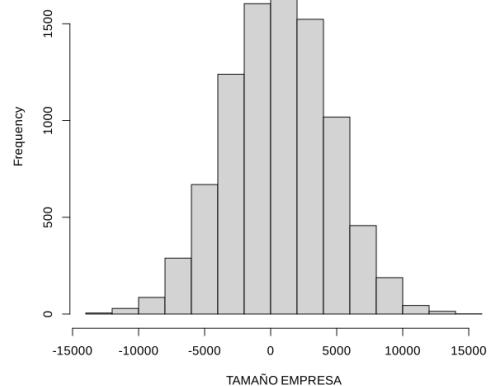
Distribución normal de salario del trabajador



Distribución normal de NIVEL DE ESTUDIOS APROBADOS



Distribución normal de TAMAÑO EMPRESA



Distribución de Poisson

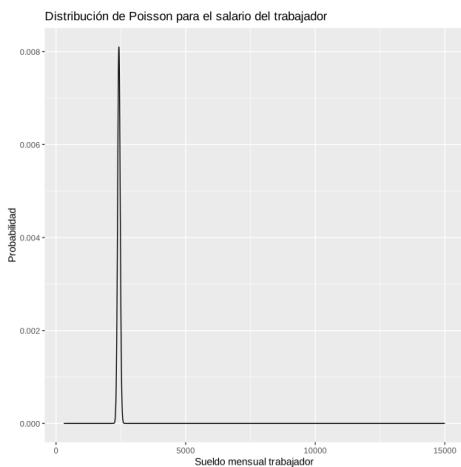


Ilustración 14. Distribución de Poisson salario del trabajador.

La distribución de Possion obtenido es de 2426.33709981168 al analizar el salario del trabajador.

Cuarto modelo construido de forma correcta por Hery

Para la elaboración del cuarto modelo, se analizaron 6 profesiones y estás van a detallarse a continuación.

En los siguientes análisis obtenidas para cada una de las profesiones, para las matrices la variable de salida o de interés es Sueldo y las variables de entrada corresponden a sexo, edad, nivel de estudios, tamaño empresa, años trabajando en la empresa ya años de recolección de los datos.

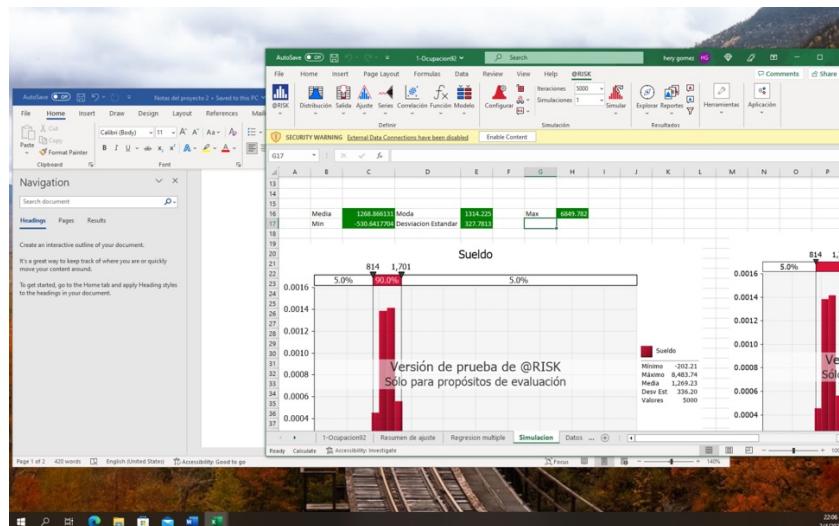


Ilustración 15. Máquina de Hery con Windows, analizando los resultados.

Peones agropecuarios

La siguiente tabla muestra la matriz de correlación de las variables seleccionadas para la construcción del modelo de los Peones agropecuarios:

Correlaci ón	Sexo	Edad	Nivel de estudios	Tamaño empresa	Años trabajad os	Sueldo	Año recolecci ón
Sexo	1.000						
Edad	-0.071	1.000					
Nivel de estudios	-0.046	-0.378	1.000				

Tamaño empresa	0.129	-0.079	0.027	1.000			
Años trabajados	-0.143	0.706	-0.328	-0.214	1.000		
Sueldo	-0.180	0.076	0.125	0.354	-0.057	1.000	
Año recolección	0.018	-0.151	0.286	-0.185	-0.066	-0.045	1.000

Tabla 6. Profesión: peones agropecuarios

Además, en la siguiente tabla se observa la distribución de probabilidad ajustada para cada una de las variables observadas. El software ajusta distintas distribuciones a los datos y calcula un índice de ajuste para cada una, utilizando el de menor valor.

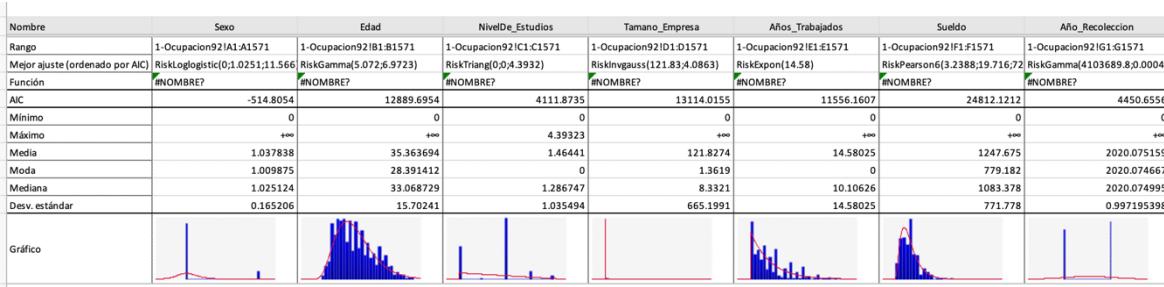


Tabla 7. Resumen de ajustes por lotes por Hery

Simulaciones:

Como siguiente paso, luego de haber obtenido las distintas correlaciones de los datos, y las distribuciones de cada conjunto de datos, se llevó a cabo la regresión múltiple. La regresión múltiple aportó los siguientes coeficientes estadísticamente significativos:

Variables	Distribuciones	Coeficientes
Sexo	1.03781451	-289.512
Edad	35.3635056	10.46832
NivelDe_Estudios	1.4644	85.9059
Tamano_Empresa	121.83	0.318053
Años_Trabajados	14.58	-12.0896
Año_Recolección	2020.082341	
Intercepto		1211.0829
Sueldo	1269.103137	

Además, también se calculo las distribuciones y la salida con @Risk utilizando la formula de la regresión con los coeficientes. El resultado de la simulación con 5000 iteraciones es el siguiente:

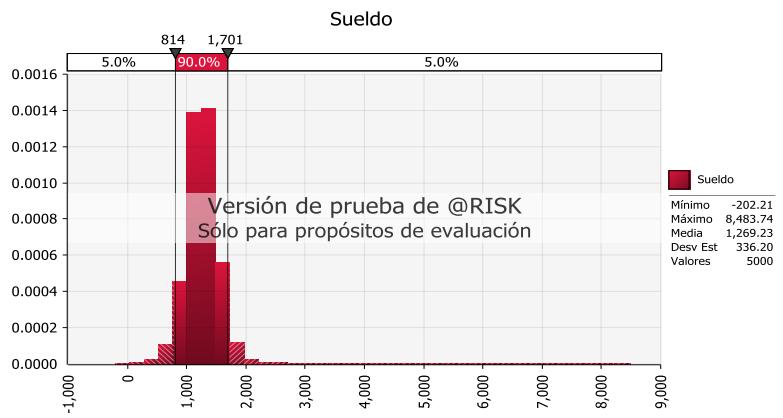


Tabla 8. Simulación peones agropecuarios por Hery.

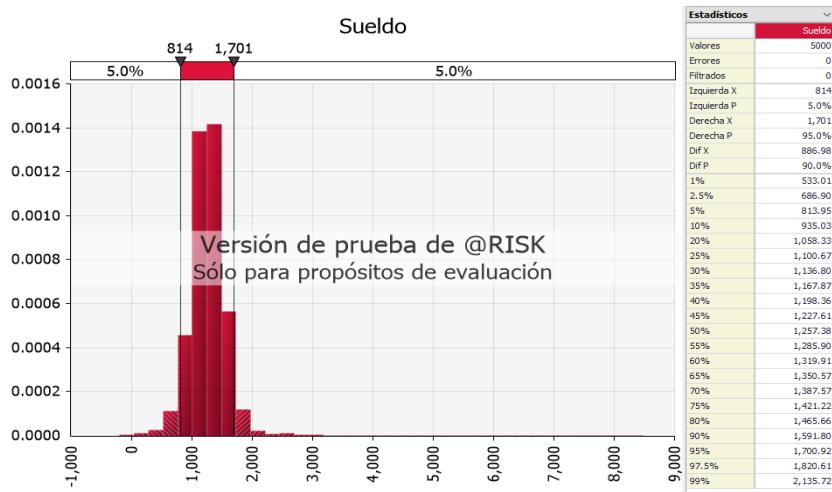


Tabla 9. Simulación datos obtenidos por Hery.

Vendedores de productos para higiene personal

Para el caso de los vendedores la matriz de correlación es la siguiente:

Correlación	Sexo	Edad	Nivel de estudios	Tamaño empresa	Sueldo	Años trabajados	Año recolección
Sexo	1.000						
Edad	0.009	1.000					
Nivel de estudios	0.041	0.007	1.000				

Tamaño empresa	-0.136	0.193	0.341	1.000			
Sueldo	-0.169	0.389	0.421	0.596	1.000		
Años trabajados	-0.114	0.573	0.019	0.163	0.338	1.000	
Año recolección	0.025	-0.005	0.057	-0.027	-0.030	-0.049	1.000

Tabla 10. Profesión: vendedores

La siguiente tabla muestra las distribuciones de probabilidad ajustadas a cada una de las variables.

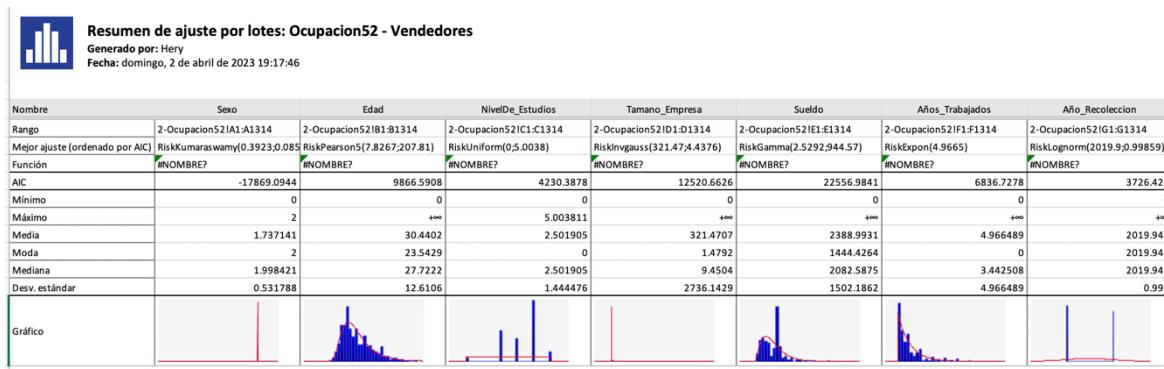


Ilustración 16. Resumen ajuste por lotes de vendedores.

Simulaciones:

Luego se realizó la regresión multiple, y se obtuvo la distribución de probabilidad para cada una de las variables con @Risk. El resultado es el siguiente:

Variables	Distribuciones	Coeficientes
Sexo	1.737139938	-449.645
Edad	30.44076933	35.54426
NivelDe_Estudios	2.5019	566.9604
Tamaño_Empresa	321.47	0.084749
Años_Trabajados	4.9665	28.35657
Intercepto		-82.3173
Sueldo	1805.136423	

Se puede observar cada uno de los coeficientes, y la salida de la simulación con 5,000 iteraciones utilizando la fórmula de la regresión. El resultado de la simulación es el siguiente:

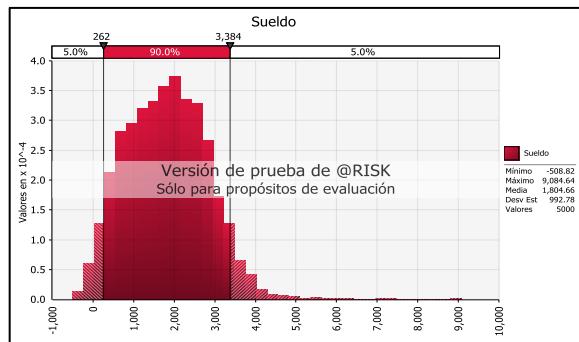


Ilustración 17. Simulación.

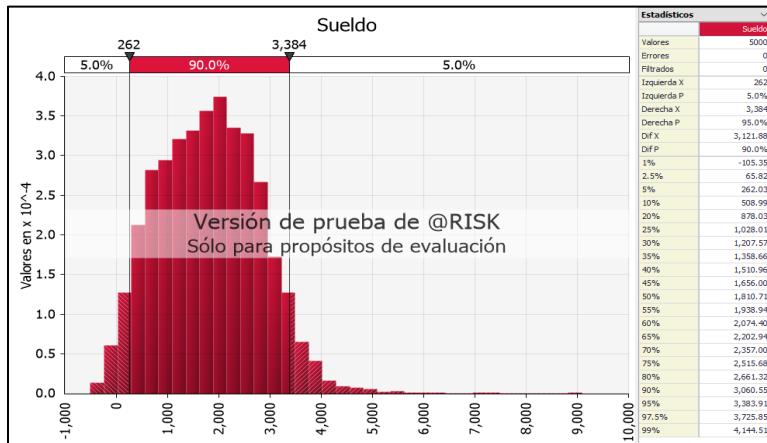


Ilustración 18. Resultados de simulación.

Personal de Limpieza

La siguiente tabla muestra la matriz de correlación:

Correlación	Sexo	Edad	Nivel de estudios	Tamaño empresa	Sueldo	Años trabajados	Año recolección
Sexo	1.000						
Edad	-0.008	1.000					
Nivel de estudios	-0.085	-0.319	1.000				
Tamaño empresa	-0.403	-0.005	0.085	1.000			
Sueldo	-0.284	0.190	0.175	0.331	1.000		

Años trabajados	-0.005	0.497	-0.267	0.064	0.191	1.000	
Año recolección	-0.035	-0.080	0.005	-0.039	-0.140	-0.001	1.000

Tabla 11. Profesión limpiadores.

Ahora, en la siguiente tabla se muestra las distribuciones ajustadas para cada una de las variables.

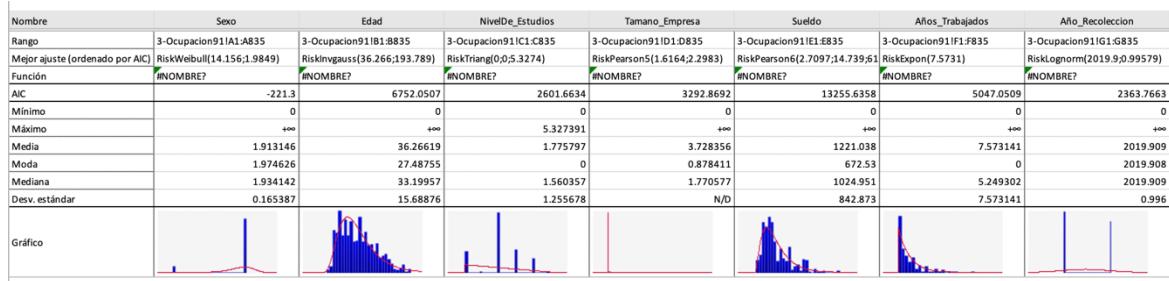


Tabla 12. Ajustes por lote generado por Hery

Simulaciones:

Para generar las simulaciones, primero se generó la regresión. También se utilizaron las fórmulas de @Risk para cada una de las distribuciones ajustadas, previamente calculadas en el apartado anterior. El resultado es el siguiente:

Variables	Distribuciones	Coeficientes
Sexo	1.913172057	-676.92
Edad	36.266	10.92758
NivelDe_Estudios	1.7758	168.6257
Tamano_Empresa	3.728585334	0.588263
Años_Trabajados	7.5731	5.717044
Intercepto		1685.992
Sueldo	1132.16184	

Luego se procedió a realizar la simulación con 5,000 iteraciones. El resultado de la simulación para la variable de salida es el siguiente:

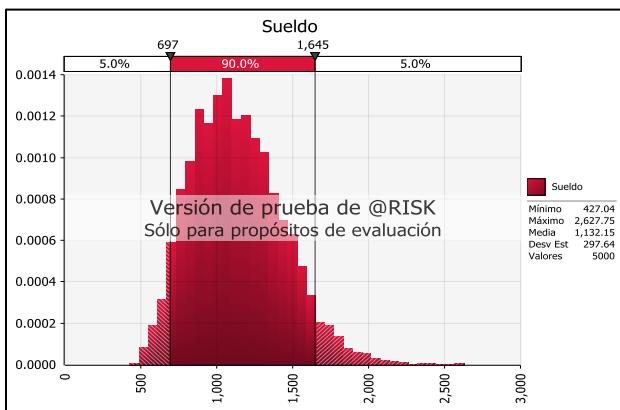


Tabla 13. Simulación.

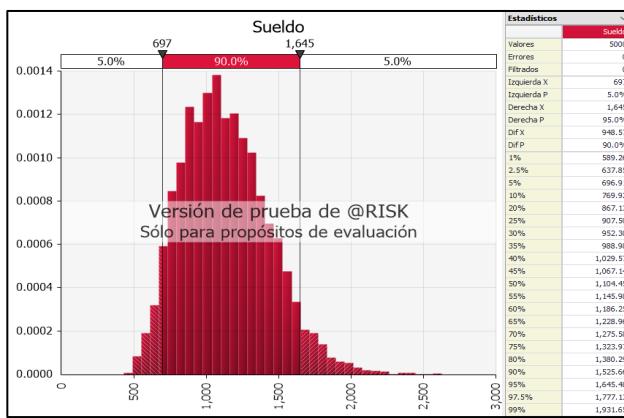


Tabla 14. Resultados de simulación.

Obreros de la construcción

La siguiente tabla corresponde a la matriz de correlación obtenida al hacer el análisis en colab.

Correlación	Sexo	Edad	Nivel de estudios	Tamaño empresa	Sueldo	Años trabajados	Año recolección
Sexo	1.000						
Edad	-0.074	1.000					
Nivel de estudios	0.096	-0.390	1.000				
Tamaño empresa	0.104	0.023	0.135	1.000			
Sueldo	-0.059	0.262	0.086	0.458	1.000		
Años trabajados	-0.114	0.706	-0.329	-0.019	0.250	1.000	

Año recolección	-0.075	0.000	-0.054	-0.006	0.004	-0.009	1.000
------------------------	--------	-------	--------	--------	-------	--------	-------

Tabla 15. Profesión: oficiales de la construcción

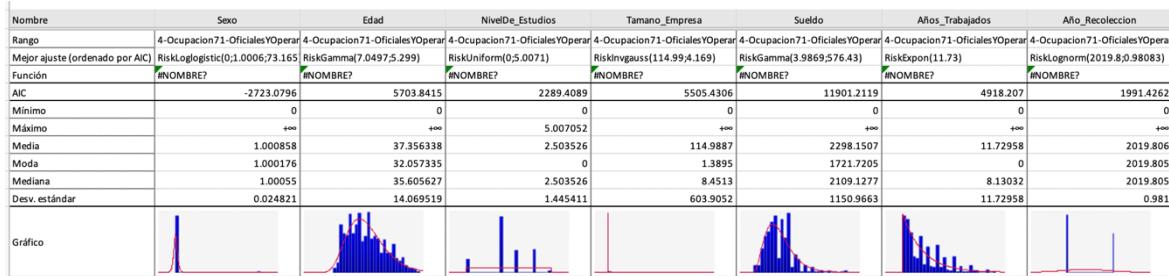


Tabla 16. Resumen de ajustes por lote por Hery

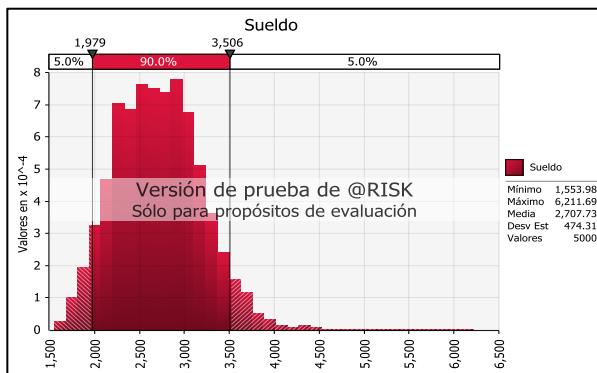


Tabla 17. Simulaciones por Hery.

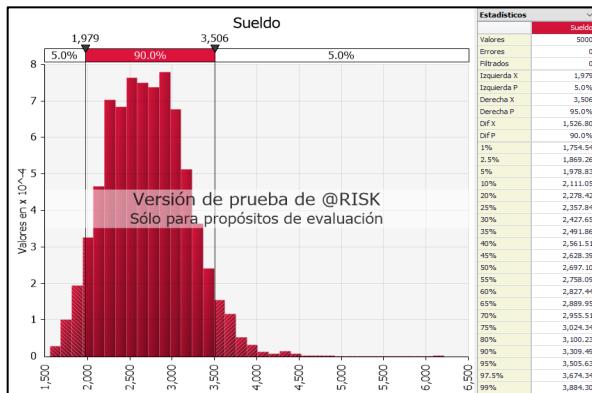


Tabla 18. Resultados de la simulación por Hery.

Profesionales de la enseñanza

La siguiente tabla corresponde a la matriz de correlación de las variables:

Correlación	Sexo	Edad	Nivel de estudios	Tamaño empresa	Sueldo	Años trabajados	Año recolección
Sexo	1.000						

Edad	-0.104	1.000					
Nivel de estudios	-0.057	-0.025	1.000				
Tamaño empresa	0.012	0.053	0.192	1.000			
Sueldo	-0.056	0.520	0.076	-0.007	1.000		
Años trabajados	-0.054	0.740	0.004	-0.032	0.585	1.000	
Año recolección	0.024	0.072	0.025	0.009	0.051	0.129	1.000

Tabla 19. Profesión: profesionales de la enseñanza.

La siguiente tabla muestra las distribuciones ajustadas para cada una de las variables.

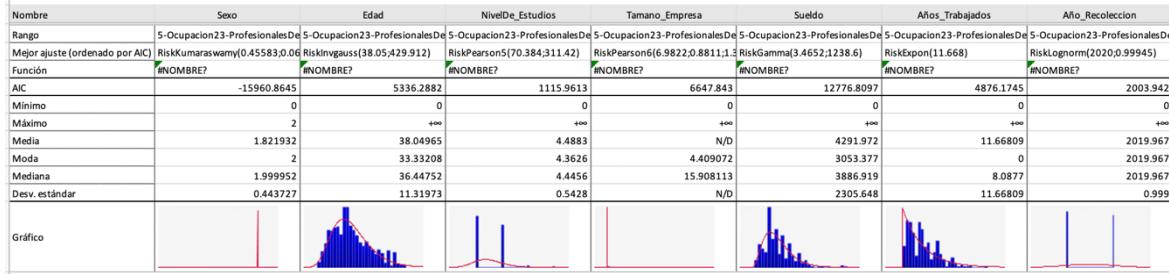


Tabla 20. Resumen de ajustes por lote por Hery

Simulaciones:

Antes de correr una simulación se realizó la regresión multiple. Los coeficientes de correlación y las variables con las funciones de distribución ajustada son las siguientes:

Variable	Distribuciones	Coeficientes
Edad	38.05	39.31686
Años_Trabajados	11.668	103.4051
Intercepto		1589.44
Sueldo		4291.97723

Luego se llevó a cabo la simulación con las 5,000 iteraciones utilizando la formula de la regresión y los coeficientes antes mostrados. El resultado de la simulación es el siguiente:

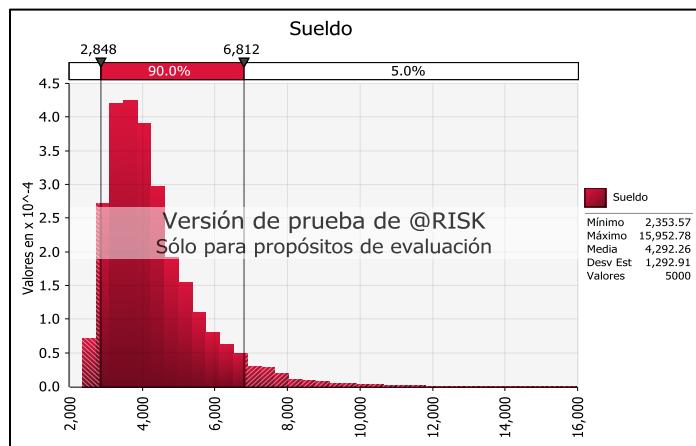


Tabla 21. Simulación de profesionales de la enseñanza por Hery

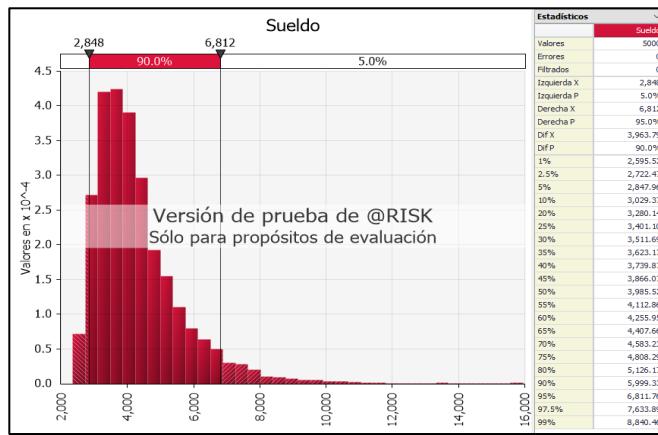


Tabla 22. Resultados de simulación por Hery.

Peones de la minería

La siguiente tabla muestra la correlación de las variables:

Correlación	Sexo	Edad	Nivel de estudios	Tamaño empresa	Sueldo	Años trabajados	Año recolección
Sexo	1.000						
Edad	0.061	1.000					
Nivel de estudios	0.188	-0.243	1.000				
Tamaño empresa	0.281	0.070	0.229	1.000			
Sueldo	0.184	0.230	0.249	0.488	1.000		
Años trabajados	0.008	0.530	-0.244	-0.025	0.095	1.000	

Año recolección	0.032	0.018	0.030	0.000	0.009	0.025	1.000
-----------------	-------	-------	-------	-------	-------	-------	-------

Tabla 23. Profesión: peones de la minería

La siguiente tabla muestra el ajuste de distribución para cada una de las variables seleccionadas.

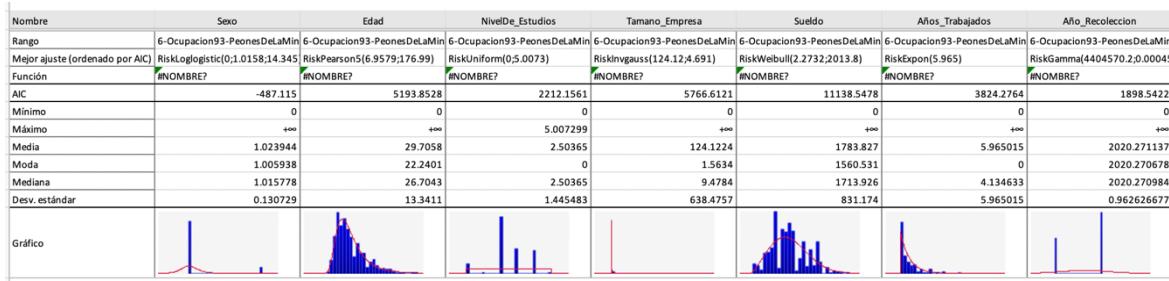


Tabla 24. Resumen de ajustes por lote por Hery

Simulaciones

Antes de realizar la simulación, se calculó la regresión multiple con las variables seleccionadas, y se aplicó la función de distribución, según la distribución ajustada obtenida en el paso anterior. El resultado es el siguiente:

Variables	Distribuciones	Coeficientes
Edad	29.70677588	19.17422
NivelDe_Estudios	2.50365	217.4936
Tamano_Empresa	124.12	0.366472
Intercepto		639.5441
Sueldo	1799.162712	

Luego se realizó la simulación con 5,000 iteraciones. El resultado para la variable de salida es el siguiente:

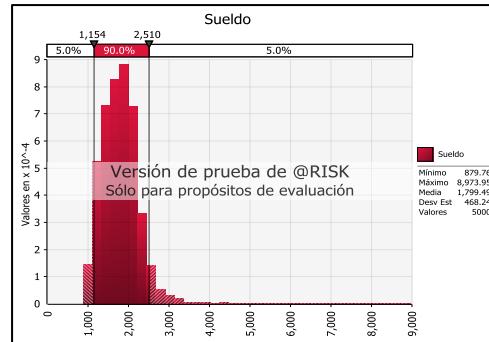


Tabla 25. Simulación por Hery

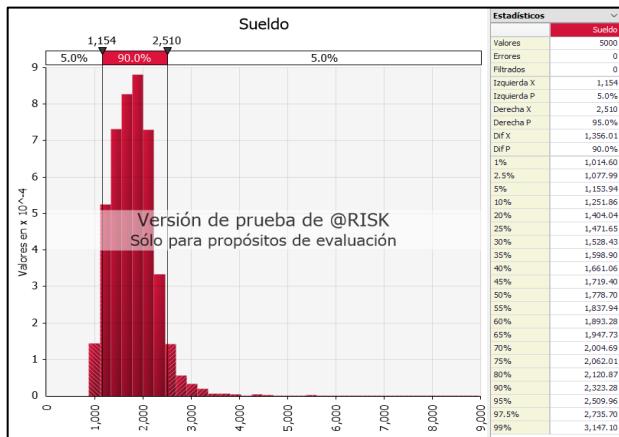


Tabla 26. Simulación datos obtenidos por Hery.

Trabajo realizado por cada estudiante

Básicamente cada uno de los integrantes trabajo de forma independiente con el objetivo de probar diferentes escenarios. Cada integrante realizó simulaciones de las profesiones por aparte, utilizando distintos softwares de simulación como Excel, @RISK, pero aplicando los mismos pasos con las restricciones de cada uno de los softwares utilizados. Luego los resultados fueron discutidos con el objetivo de utilizar el mejor enfoque o determinar posibles errores en los análisis realizados previamente por cada estudiante.

Sin embargo, un resumen de lo que cada estudió hizo:

Estudiante: Hery Fernando Gomez de Leon

- Realizo varias simulaciones de profesiones utilizando @RISK
- Descargo software @RISK
- Documentación sobre las simulaciones realizadas
- Generación del código utilizado para el análisis de los datos utilizando Colab y R.
- Investigación sobre el nombre de variables que estaban disponibles en el excel.
- Cambio el nombre de las variables que tienen ciertos códigos por nombres más detallados.
- Realización sobre la matriz de correlación
- Realización sobre datos estadísticos
- Limpieza de la base de datos, por ejemplo: no considerar filas con datos vacíos.
- Interpretación de los resultados obtenidos en @RISK

Estudiante: Santos López Tzoy

- Unión de las bases de datos del año 2019 y 2021.
- Instalación de Docker para correr Jupyter para analizar la base de datos con R.
- Realización sobre la matriz de correlación sobre otras profesiones.
- Limpieza de la base de datos, algunas filas afectaban el resultado de la matriz de correlación.

- Análisis de datos estadísticos realizados en R con Jupyter como entorno de trabajo.
- Creación de la matriz de correlación utilizando R en Jupyter.

Descargar software de Simulación @RISK

Para analizar la simulación es necesario descargar algún software de simulación, en nuestro caso, se utilizó @RISK y está puede descargarse de la página: <https://www.palisade.com/trials/>

Para obtener una copia del software es necesario completar los datos del formulario.

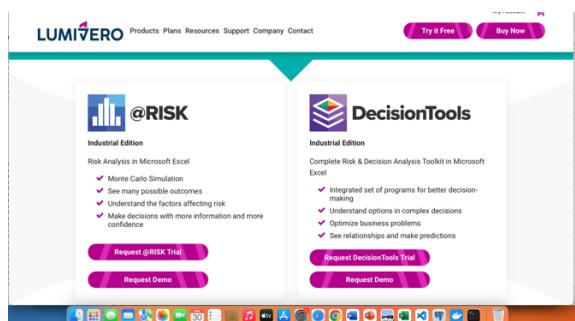


Ilustración 19. Sitio @Risk

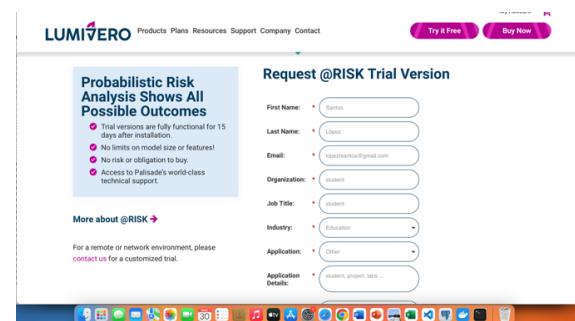


Ilustración 20. Formulario @Risk

Al completar todo el formulario y enviarlo, uno espera para recibir un enlace para obtener una copia del software.

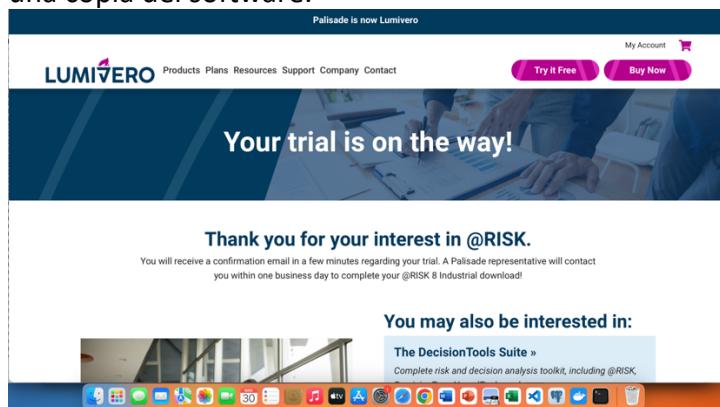
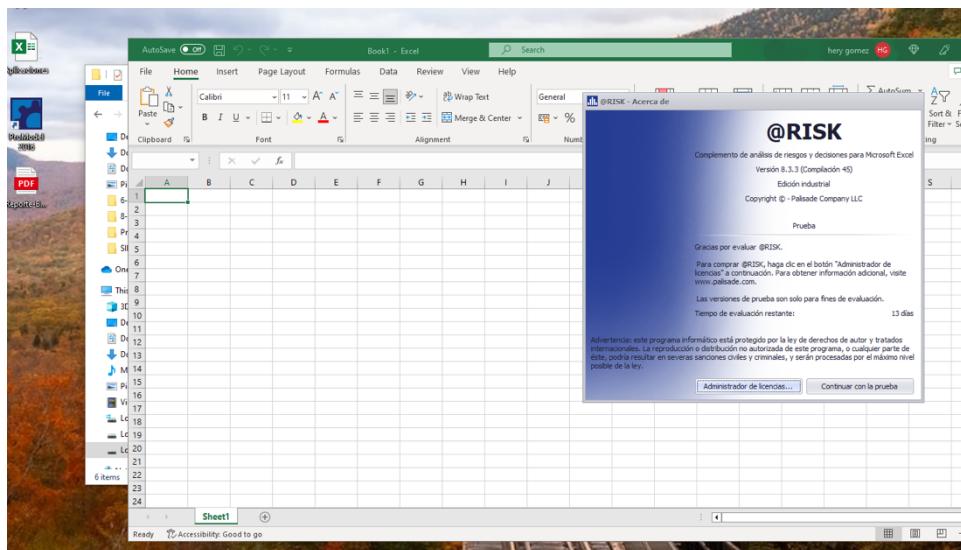


Ilustración 21. Correo @Risk

Para inicializar @Risk se debe buscar la aplicación en el botón de inicio de Windows y luego ejecutarla. Aparecerá una ventana con los detalles de la licencia, y también se abrirá el programa de Microsoft Excel.



Conclusiones

En el estudio de la elaboración del modelo utilizado para el proyecto “Salario devengado por el tipo de nivel académico” se concluye que:

- Aunque las variables de entrada pueden ser todas las que uno quiere, para el estudio solo debe tomarse en cuenta las que aporte más valor a la variable de interés (variable de salida), por ejemplo: ¿Qué tanto influye que una persona al tener más edad y un mejor nivel académico pueda obtener un mejor salario?
- La base de datos utilizado en el proyecto tenía muchas filas con valores vacíos, para el estudio es importante hacer una limpieza para no tomar en cuenta estos registros en la matriz de correlación, el inconveniente de utilizar valores vacíos es que puede provocar que la matriz de correlación sea equivocada.
- Todos los modelos construidos en el proyecto fueron documentados para tener presente los inconvenientes en las que se enfrentó en el estudio, un mal análisis puede provocar que se empieza mal al hacer el primer clic, como lo dice el libro: Tips 4 Simulations”

Referencias bibliográficas

Encuesta nacional de Empleo e Ingresos – Instituto Nacional de Estadística. **Encuesta nacional de Empleo e Ingresos – Instituto Nacional de Estadística.** 2023. Consultado en: <https://www.ine.gob.gt/encuesta-nacional-de-empleo-e-ingresos/> Disponible el: 03 de marzo de 2023

Sokolowski, J. A., & Banks, C. M. (Eds.). (2010). ***Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains***. Hoboken, NJ: John Wiley & Sons.