Examen final

150344 Jerónimo Aranda Barois

Veamos en la siguiente linea el tamaño de nuestra muestra:

Out[22]:

```
(2500, 77)
```

Es decir, nuestra base tiene 2500 observaciones de 77 columnas, ahora procedamos a hacer de esta muestra, 2 submuestras, una para los estados donde transitará el tren maya y el complemento,veamos sus tamaños y sus primeros 5 elementos.

```
(222, 78)
```

Out[31]:

	ENTIDAD	NOM_ENT	MUNICIPIO	NOM_MUN	PEA	PEA_M	PEA_F	POCUPADA	P
4338	4	CAMPECHE	4	CHAMPOTON	550	431	119	543	
41365	7	CHIAPAS	90	TAPACHULA	363	184	179	351	
10481	4	CAMPECHE	10	CANDELARIA	272	255	17	266	
23990	31	YUCATAN	56	OXKUTZCAB	503	333	170	493	
41567	7	CHIAPAS	102	TUXTLA GUTIERREZ	2048	1322	726	1997	

5 rows × 78 columns

```
(2286, 77)
```

Out[4]:

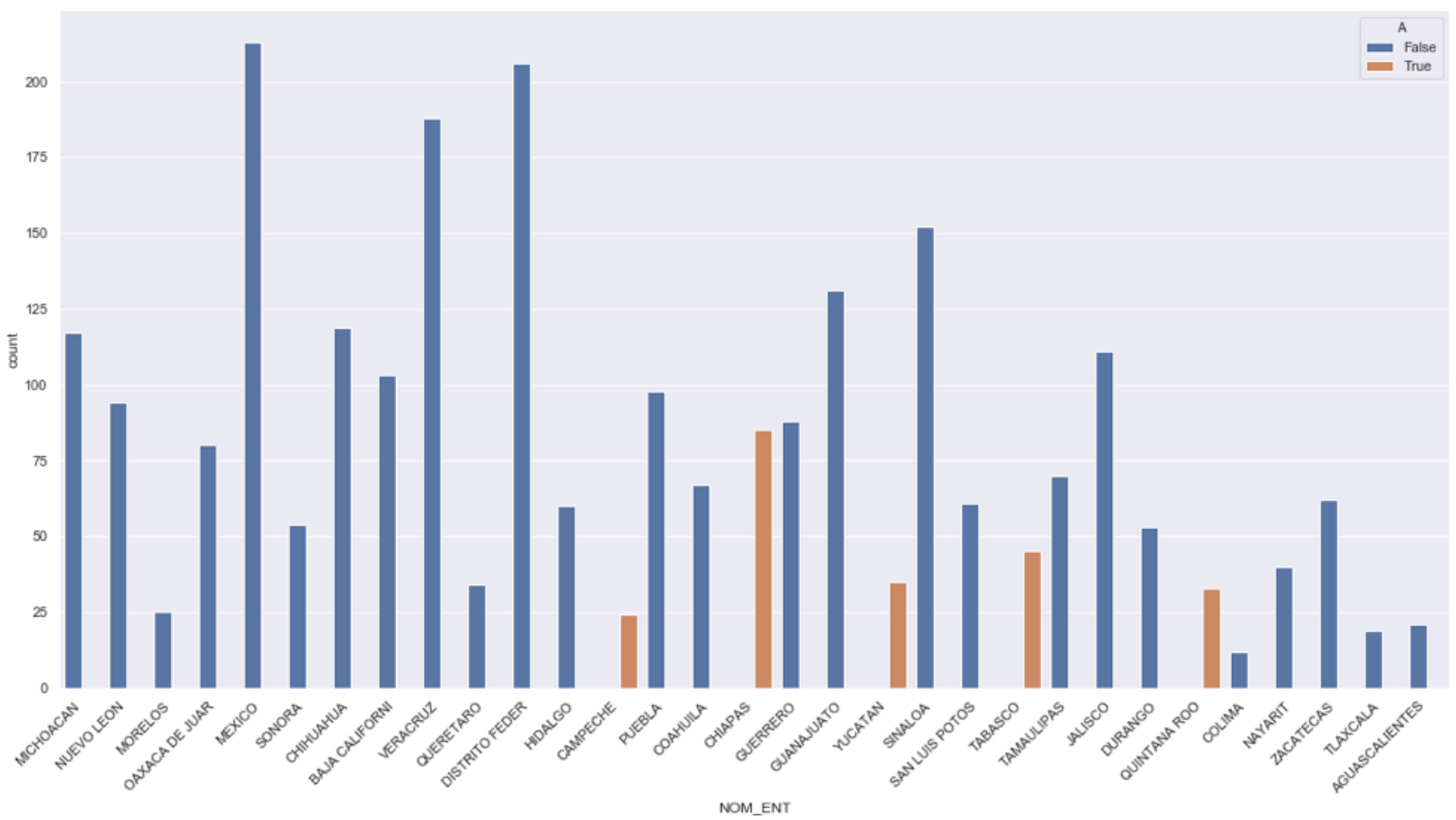
	ENTIDAD	NOM_ENT	MUNICIPIO	NOM_MUN	PEA	PEA_M	PEA_F	POCUPADA	PO
50569	9	DISTRITO FEDER	6	IZTACALCO	633	390	243	604	
31592	14	JALISCO	18	LA BARCA	548	343	205	526	
25433	5	COAHUILA	35	TORREON	886	563	323	765	
8320	25	SINALOA	2	ANGOSTURA	57	52	5	56	
2615	24	SAN LUIS POTOS	2	ALAQUINES	45	39	6	31	

5 rows × 77 columns

La división parece no corresponder al número de estados que existen en la república, talvez corresponde al número de distritos que existen, esperemos no afecten los resultados.

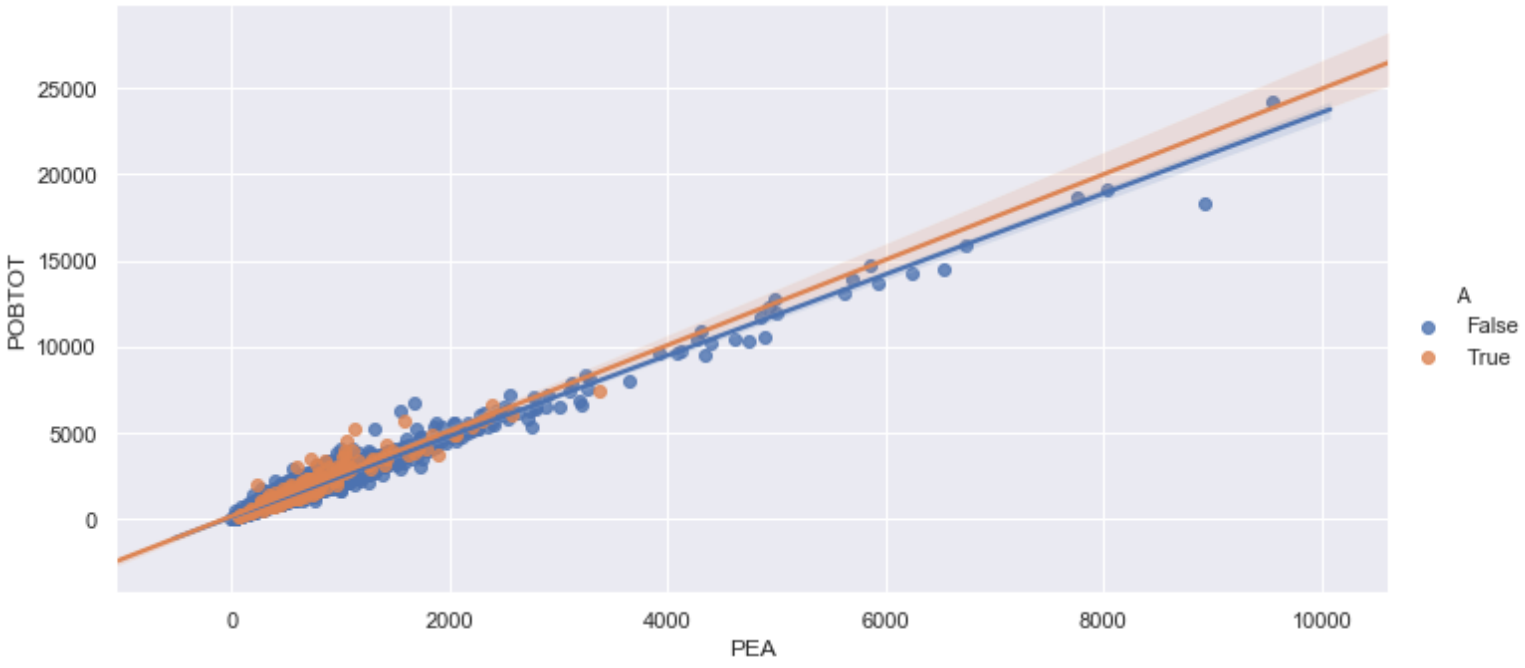
Pregunta 1

Para empezar veamos como están distribuidas nuestras observaciones por estado y además a qué muestra pertenecen.



En general nuestra submuestra de interés A tiene menos observaciones que las de algunos otros municipios pero también más que algunos otros, esto puede ser respecto a las distritaciones del país.

A continuación veamos como es la Población económicamente activa respecto a la población total.

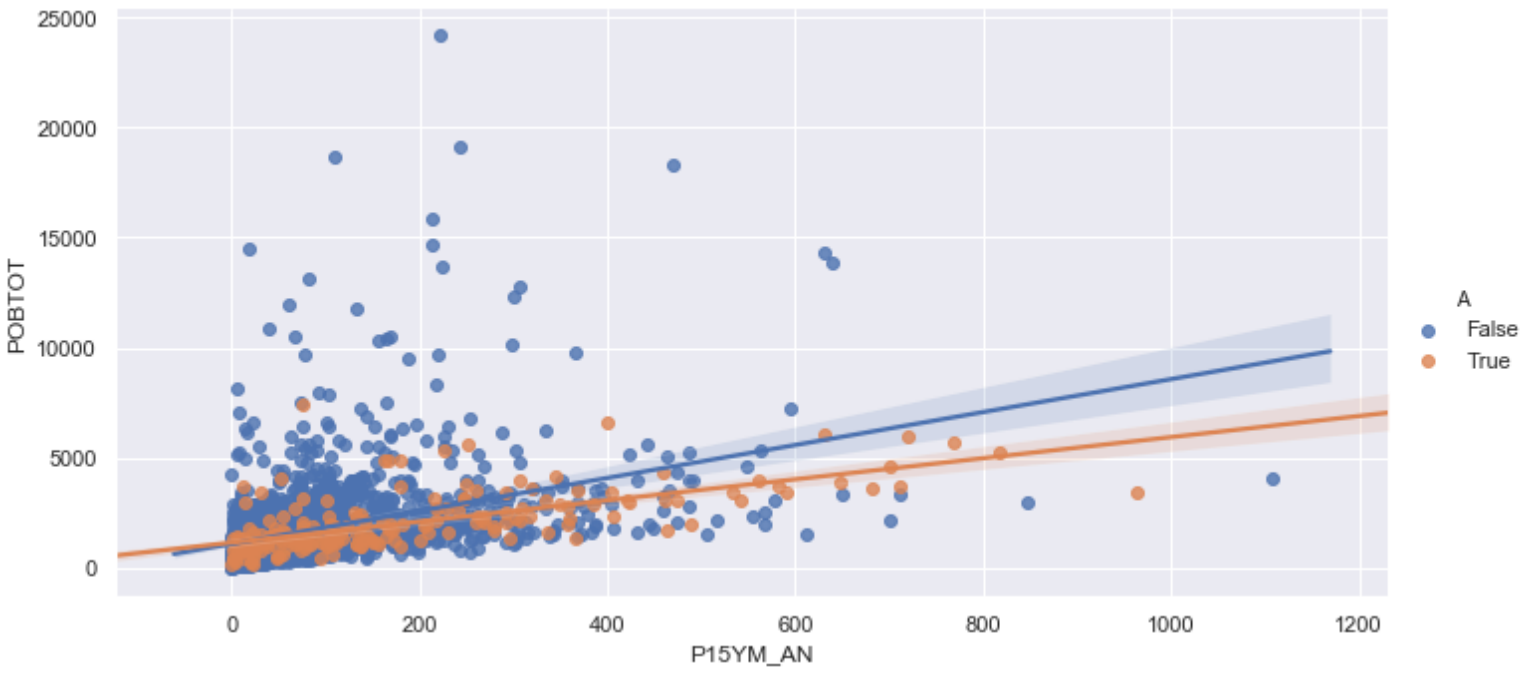


Aunque nuestros estados elegidos 'A' parecen estar entre el bonche, parecen tener un poco menos de población económicamente activa frente a su población total que otros estados. Para soportar esta hipótesis habría que hacer una regresión lineal y ver que las β's sean significativas y distintas, es decir la prueba de hipótesis correspondientes.

Ahora de nuevo veamos nuestra población total frente a la población de 15 años o más que es analfabeta.

Out[67]:

```
<seaborn.axisgrid.FacetGrid at 0x1a3a66bc88>
```

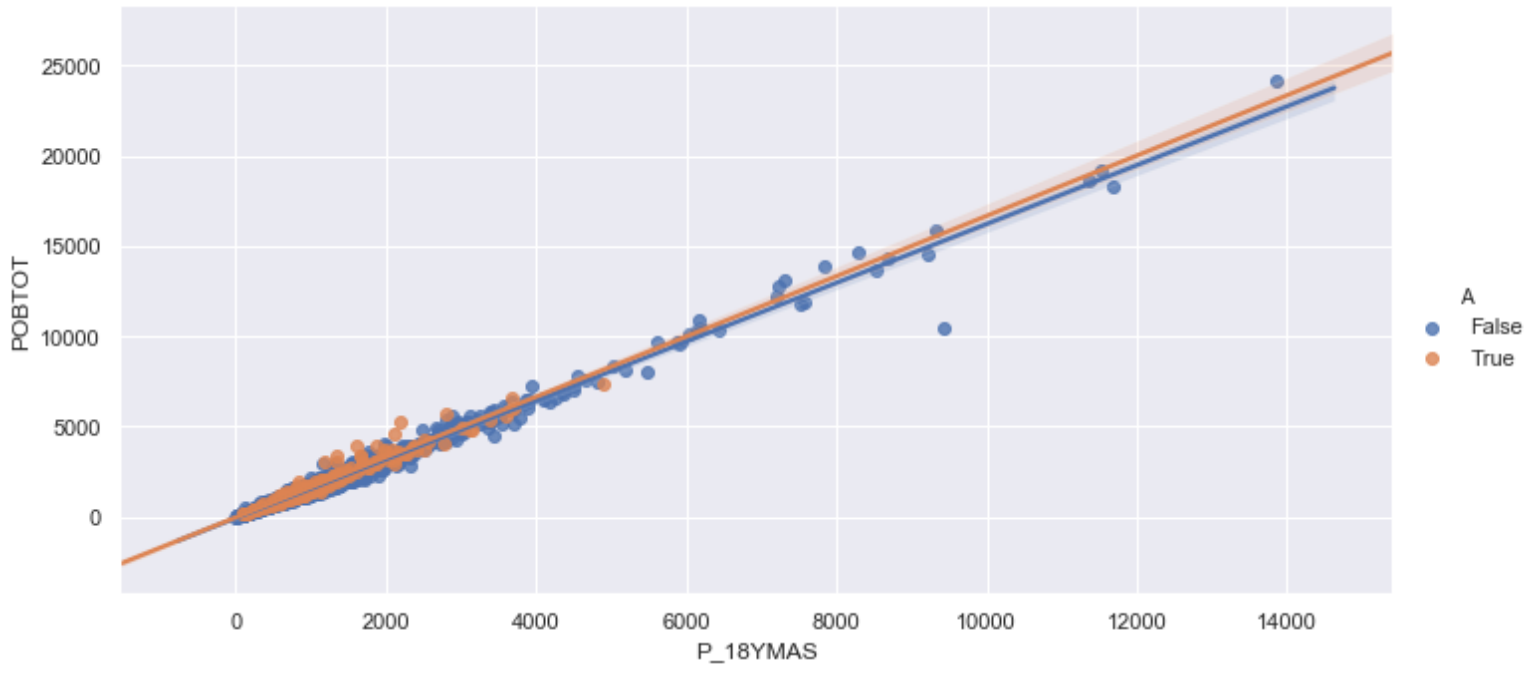


En este caso parece ser inversa la relación es decir, si bien en ambos grupos hay un poco de casi todos los ratios, parece ser consistente que en nuestro grupo A, parece haber una población de 15 años más analfabeta. De nuevo habría que respaldar estas hipótesis con las correspondientes pruebas. Sin embargo, estas gráficas pueden ser suficientes pues el paquete ya ajusta los modelos lineales.

Veamos como es la población de 18 y mas frente a la población total.

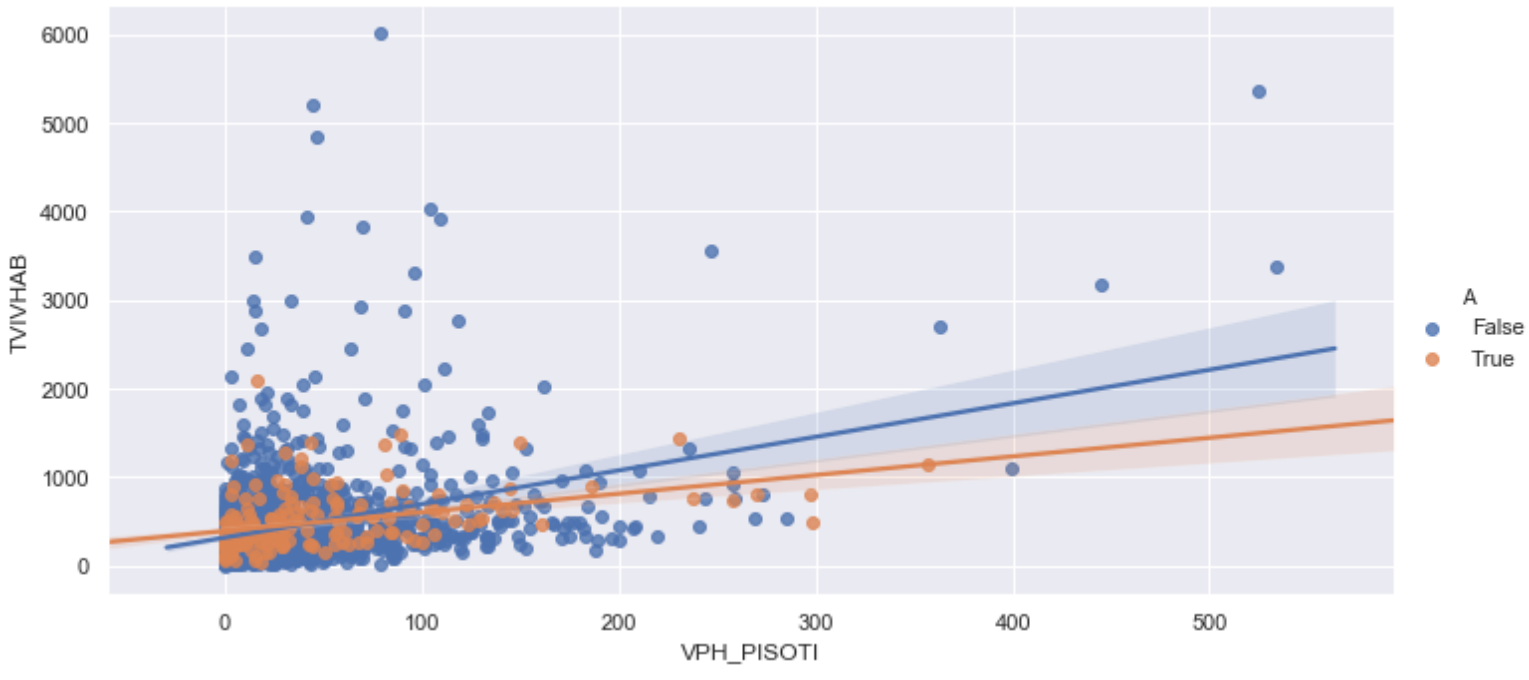
Out[68]:

```
<seaborn.axisgrid.FacetGrid at 0x1a3a886d30>
```



Las poblaciones parecen tener casi la misma tasa $P18/Ptot$.

Ahora veamos viviendas viviendas con piso de tierra frente a viviendas habitadas

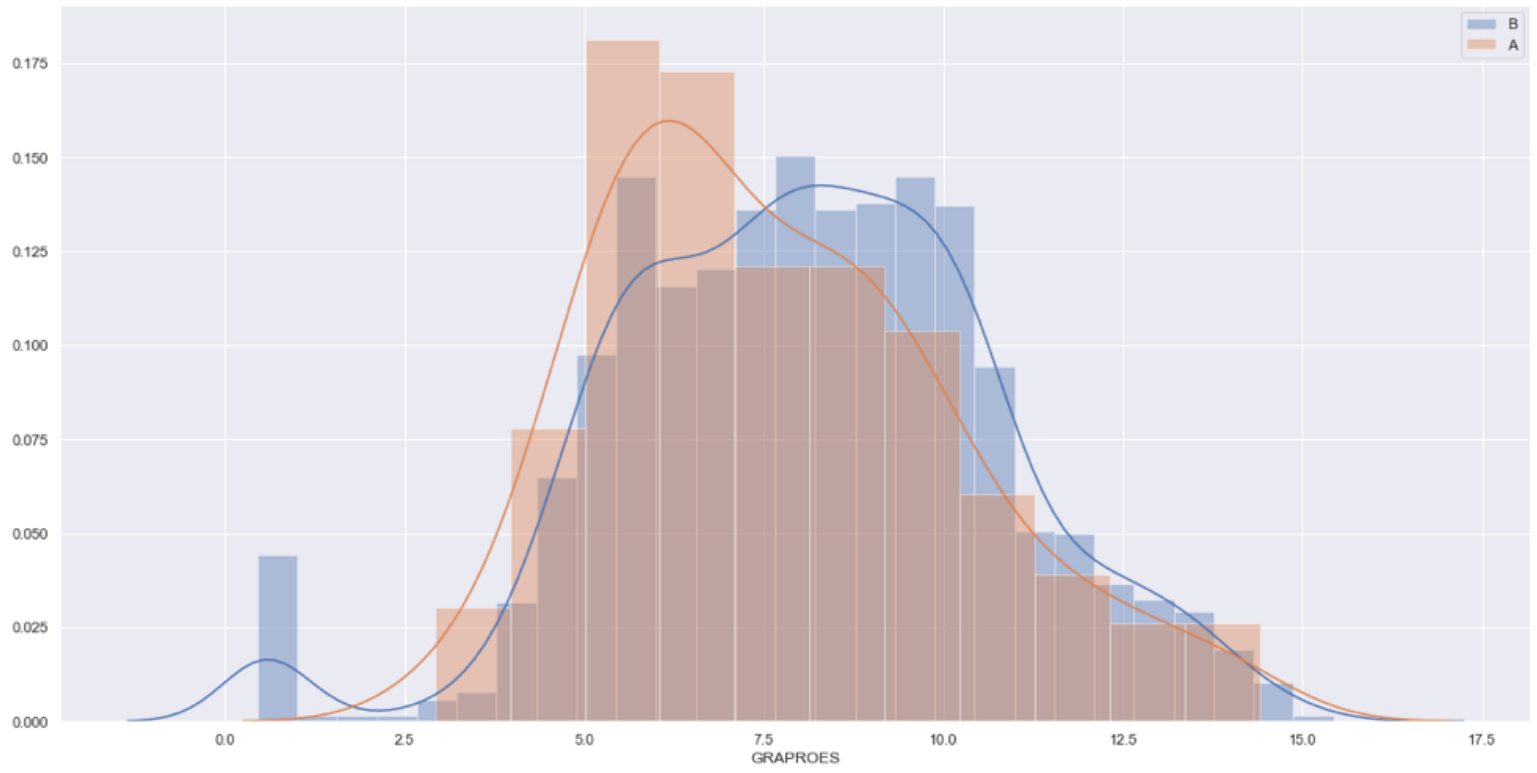


Parece haber una diferencia significativa en este ratio.

Veamos cómo son las distribuciones de nuestra población respecto a sus grados promedio de escolaridad.

Out[44]:

```
<matplotlib.legend.Legend at 0x1a3b110198>
```



Parecen ser distribuciones muy distintas, podemos concluir por esta figura y todas las anteriores que nuestra muestra parece tener peores condiciones económicas, esto aunque incurre en mayor riesgo de inversión, también significa mayores retornos esperados. Sin embargo no se puede esperar una industria muy especializada, sino que más bien una que mejor aproveche el entorno. Algo como la maquilación pues la mano de obra debe ser más barata.

En otras palabras, y suponiendo que las condiciones económicas peores significan mayores areas de oportunidad yo haría las siguientes recomendaciones.

- La participación en la economía es peor en nuestra submuestra A.
- La educación es peor en nuestra submuestra A, lo que no soportaría industria tan especializada.
- Parecen tener una distribución demógrafica parecida ambos grupos.
- La calidad de la vivienda es más baja en el grupo A.

Muchos **retos** pero grandes **areas de oportunidad**.

Pregunta 2

Para nuestro PCA utilizaremos las siguientes variables:

```
['ENTIDAD', 'MUNICIPIO', 'PEA', 'PEA_M', 'PEA_F', 'POCUPADA', 'POCUP
ADA_M', 'POCUPADA_F', 'PDESOCUP', 'PDESOCUP_M', 'PDESOCUP_F', 'P15YM
_AN', 'P15YM_AN_M', 'P15YM_AN_F', 'P15YM_SE', 'P15YM_SE_M', 'P15YM_S
E_F', 'P15PRI_IN', 'P15SEC_CO', 'P15SEC_COM', 'P15SEC_COF', 'GRAPROE
S', 'GRAPROES_M', 'GRAPROES_F', 'VPH_CEL', 'POBTOT', 'POBMAS', 'POBF
EM', 'P_18YMAS', 'P_18YMAS_M', 'P_18YMAS_F', 'P_18A24', 'P_18A24_M',
'P_18A24_F', 'P_15A49_F', 'P_60YMAS', 'P_60YMAS_M', 'P_60YMAS_F', 'R
EL_H_M', 'POB0_14', 'POB15_64', 'POB65_MAS', 'P3YM_HLI', 'P3YM_HLI_M
', 'P3YM_HLI_F', 'P3HLINHE', 'P3HLINHE_M', 'P3HLINHE_F', 'PSINDER',
'PDER_SS', 'VIVTOT', 'TVIVHAB', 'TVIVPAR', 'VIVPAR_HAB', 'TVIVPARHAB
', 'PROM_OCUP', 'PRO_OCUP_C', 'VPH_PISOTI', 'VPH_1DOR', 'VPH_2YMASD'
, 'VPH_1CUART', 'VPH_2CUART', 'VPH_3YMASC', 'VPH_C_ELEC', 'VPH_S_ELE
C', 'VPH_AGUADV', 'VPH_AGUAFV', 'VPH_EXCSA', 'VPH_DRENAJ', 'VPH_NODR
EN', 'VPH_C_SERV', 'VPH_RADIO', 'VPH_TV', 'VPH_REFRI', 'CONSECUTIVO'
, 'A']
```

Con una cantidad de variables total:

Out[33]:

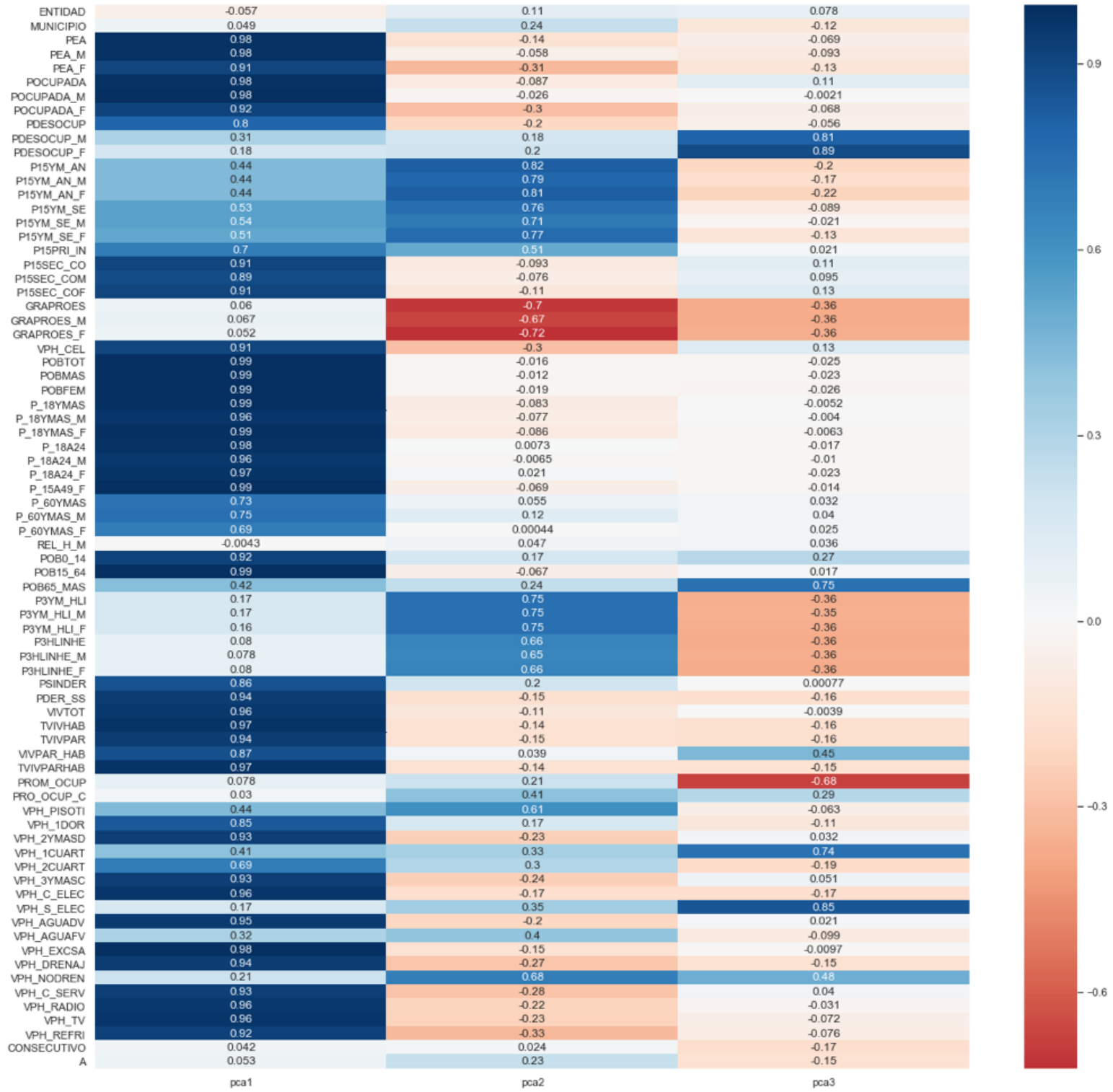
```
76
```

Apliquemos **PCA** y veamos la explicación de las primeras 3 variables:

Out[34]:

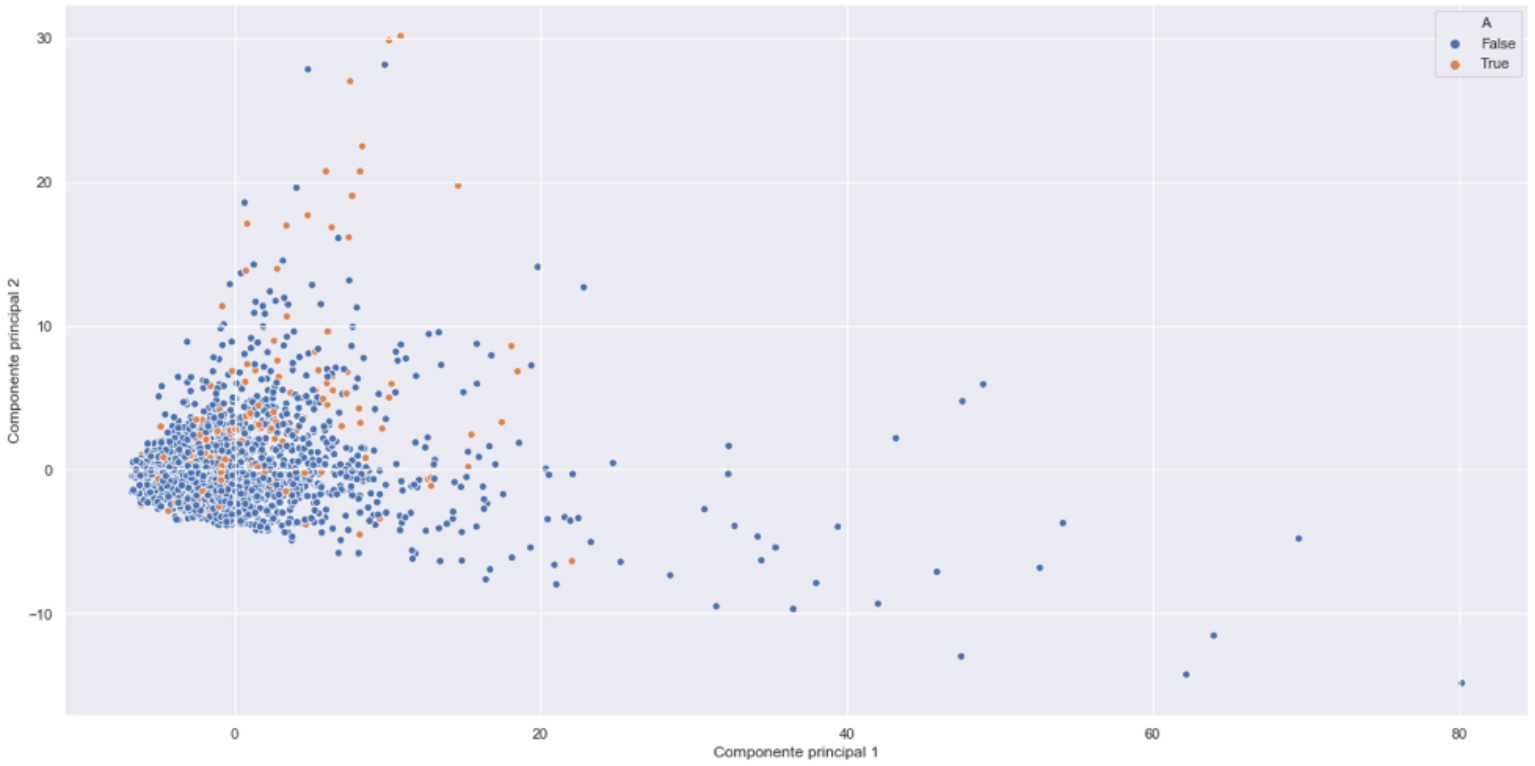
```
array([41.58417124, 11.28105276,  6.04032318])
```

Veamos nuestras correlaciones con las variables. Es interesante ver como las variables importantes en algunos componentes dejan de ser las variables importantes en el resto.

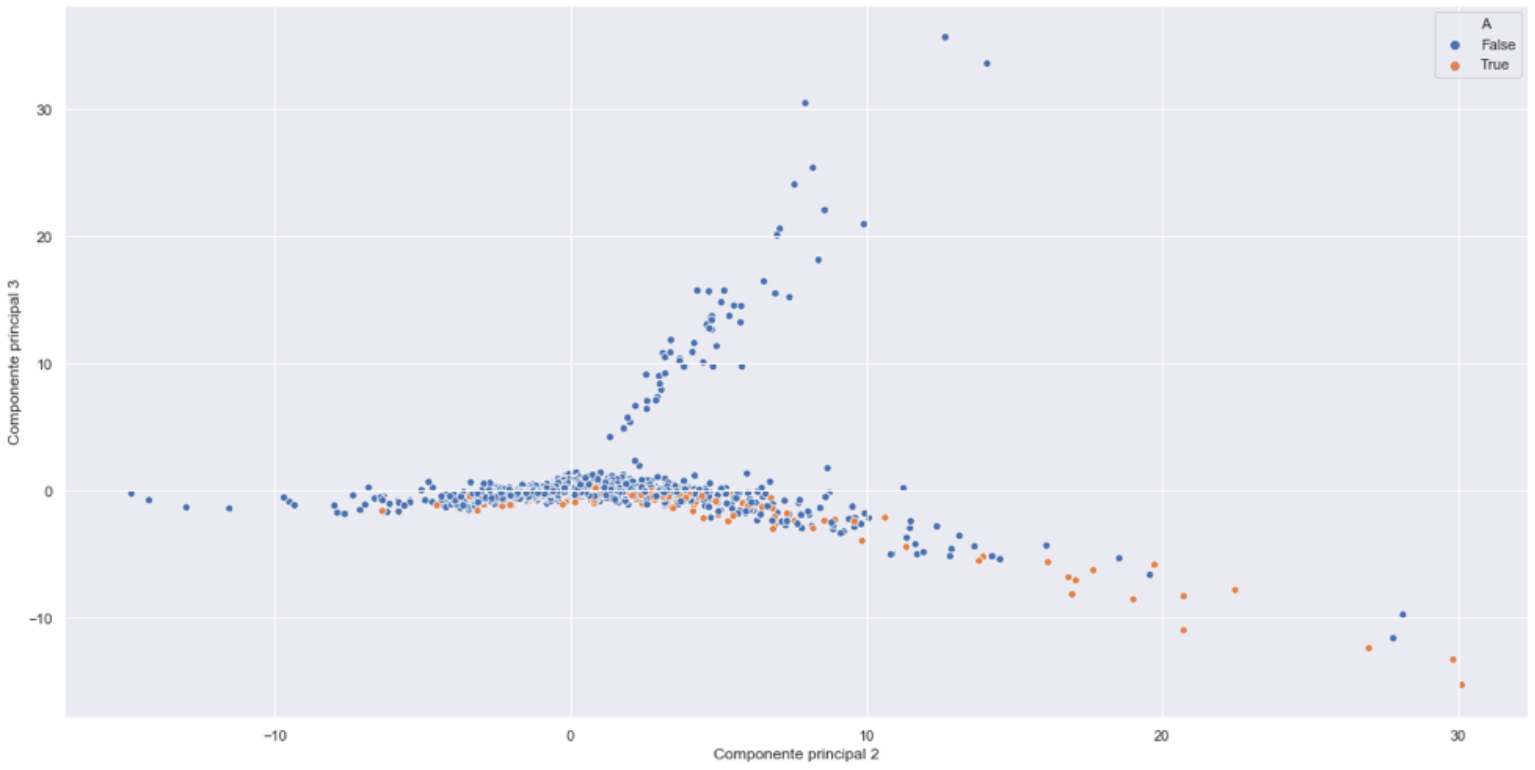


Interpretar estas 76 variables puede ser complicado, sin embargo concentrémonos en las últimas 2 columnas, es fácil ver que nuestra 2da componente está muy relacionada con poblaciones de 15 años analfabetas y sin escolaridad y negativamente correlacionado con el grado promedio de estudios. Es decir nuestra segunda componente engloba mucha de la información de la escolaridad. La tercera componente está muy negativamente correlacionada con el promedio de ocupantes de viviendas particulares ocupadas y positivamente con las viviendas sin electricidad, de un solo cuarto y con la desocupación de hombres y mujeres.

A continuación veamos nuestras 2 primeras componentes:



No parecen separar bien a nuestras muestras sin embargo las segundas 2 componentes parecen ser mucho más explicativas para la región. Veamos estas 2:



Es interesante ver como las componentes principales 2 y 3 separan mejor a nuestra muestra. La 2 especificamente capturando la información de escolaridad. Es decir **negativamente** correlacionada con la escolaridad. Y la 3 capturando información de desocupación laboral **positivamente** y ocupación de viviendas **negativamente**.

Pregunta 3

Por la figura anterior es fácil ver que la proyección en las componentes 2 y 3 podría ser un buen criterio de separación, sin embargo intentemos el **LDA** entrenado con la muestra de **2500** y probado con la base de datos **real**. Veamos su **matriz de confusión**:

Que tiene la siguiente **precisión**:

```
Out[69]:
```

```
0.9481663744456396
```

Es decir una **Precisión** bastante buena.

Pregunta 4

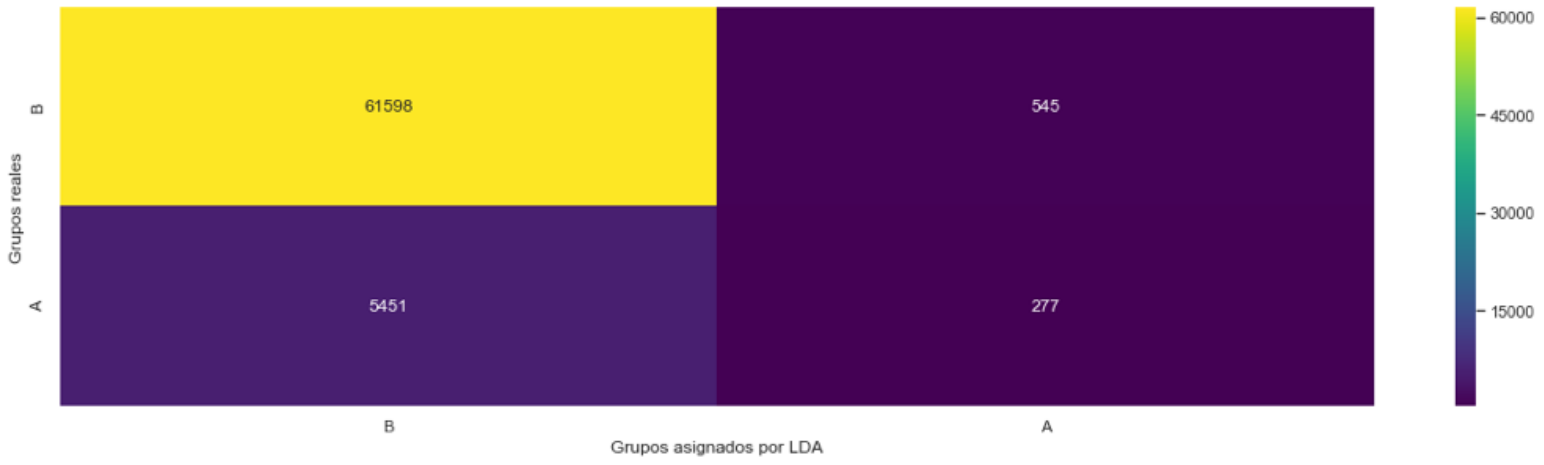
Es claro, como se veía desde la pregunta 1, que las poblaciones A y B son bastante distintas, pues el modelo **LDA** fue entrenado con 2500 de las más de 67000 observaciones y tiene una precisión del **95%** en el total de estas. Esto quiere decir que el conjunto de variables que usamos 67, separan bien a nuestras poblaciones. Sin embargo, podríamos mejorar este modelo e intentar ver como disminuye la precisión si es que solamente utilizamos alguna muestra de entre el total de nuestras 67 variables.

A continuación se intentará con las siguientes variables por elección con base en los PCA's anteriores:

```
['GRAPROES', 'PROM_OCUP', 'VPH_PISOTI', 'PDESOCUP_F', 'PDESOCUP_M',
'VPH_S_ELEC', 'VPH_1CUART', 'VPH_NODREN', 'POB65_MAS']
```

Veamos matriz de confusión y precisión:

```
0.9116559355247454
```



Si bien vemos que la precisíon es **alta**, esto está dado porque nuestras poblaciones A y B están desbalanceadas. Es decir nuestro grupo B es muy grande, mientras que nuestro grupo A es muy pequeño. Alguna mejor medida que la precisión ayudaría a ponderar mejor el número de falsos positivos y falsos negativos.