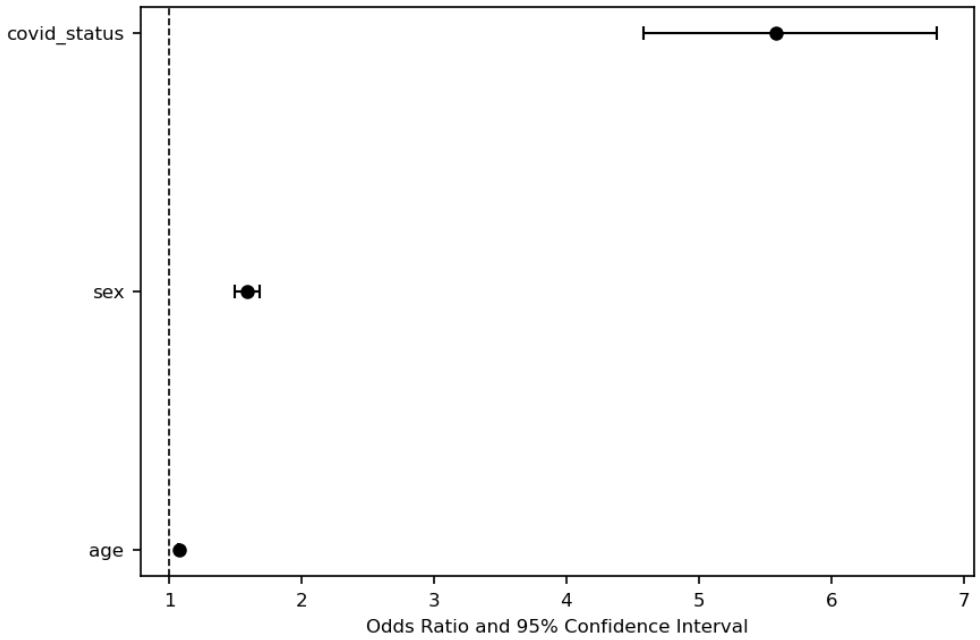
Figures

	Odds Ratio	p-value	5%	95%	significant
covid_status	2.841336	8.996275e-32	2.386385	3.383022	True
sex	1.807317	0.000000e+00	1.765253	1.850383	True
age	1.071549	0.000000e+00	1.070798	1.072301	True

Logit Regression Results

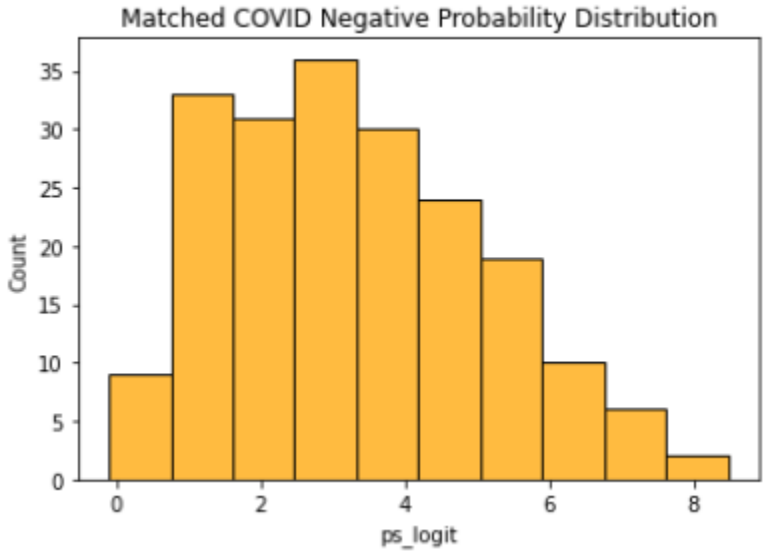
Dep. Variable:	outcome	No. Observations:	50791
Model:	Logit	Df Residuals:	50787
Method:	MLE	Df Model:	3
Date:	Tue, 13 Dec 2022	Pseudo R-squ.:	0.1730
Time:	23:48:19	Log-Likelihood:	-15755.
converged:	True	LL-Null:	-19051.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.6295	0.127	-59.930	0.000	-7.879	-7.380
covid_status	1.7193	0.100	17.128	0.000	1.523	1.916
age	0.0705	0.001	71.191	0.000	0.069	0.072
sex	0.4615	0.030	15.232	0.000	0.402	0.521

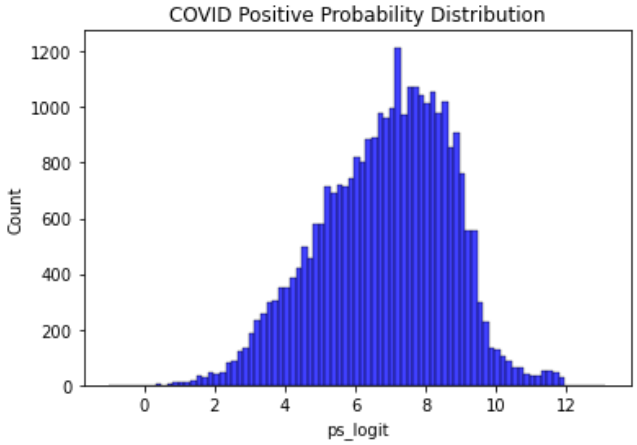


These figures show the odds ratio results (with age and sex as the confounding variables). The first figure is a compilation of the second and third - the important components are that the odds ratio numbers were significant for all of the variables (with a p-value less than 0.001 for all three variables). I observed that age and sex were indeed contributing factors to the differences in

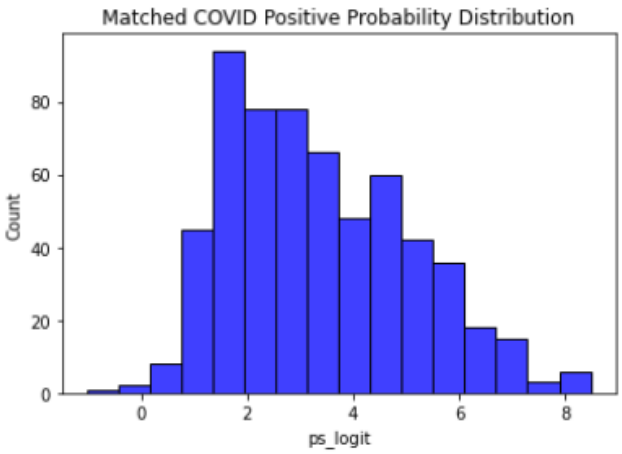
mortality between the datasets (older individuals and males were more likely to die from COVID), but even after these variables were accounted for, COVID+ patients were still more likely to die than COVID- patients (odds ratio of 2.84).



This figure shows the COVID- probability distribution for the propensity matching. The logistic regression model created this distribution of propensity scores which was used to create the matched cohort.



This figure shows the COVID+ probability distribution before matching occurred. As we can see, the shape of the distribution is not similar to the shape of the COVID- distribution, which reflects inherent biases in the datasets that may influence the dependent variable (in this case - mortality).



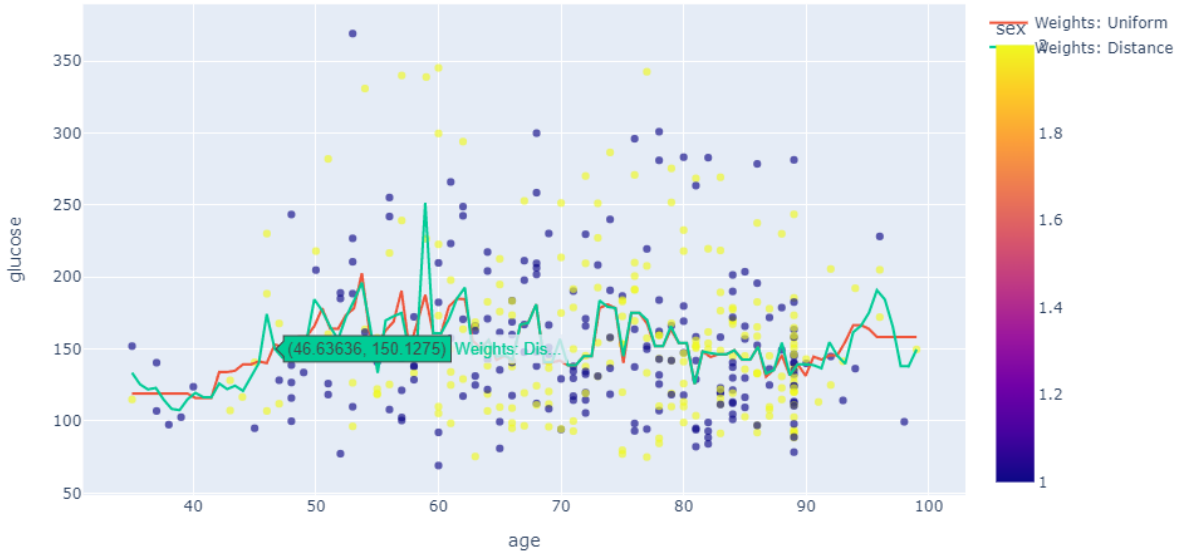
This figure shows the COVID+ probability distribution for the propensity matching. The propensity matching was successful since the distribution of this matched cohort was similar to the shape of the COVID- distribution.

	noncovid	covid	p-value	significant
Age - Mean	74.047619	42.601102	<0.001	True
Age - Standard Deviation	13.437241	16.651157	-	-
Male	618 (52.55%)	285830 (50.64%)	0.20	False
Diabetes	495 (42.09%)	493638 (87.46%)	<0.001	True
Hypertension	844 (71.77%)	472008 (83.63%)	<0.001	True
Mortality	159 (13.52%)	35888 (6.36%)	<0.001	True

This figure shows the t-test (for age) and chi-squared (for sex, diabetes, hypertension, and mortality) results between the COVID- cohort and the unmatched COVID+ cohort. We see that aside from sex, there is a significant difference in all of the other variables, indicating that the datasets cannot be effectively compared without adjusting for these other variables in some way.

	noncovid	matched_covid	p-value	significant
Age - Mean	74.047619	73.348333	0.30	False
Age - Standard Deviation	13.437241	13.39176	-	-
Male	618 (52.55%)	291.0 (48.5%)	0.12	False
Diabetes	495 (42.09%)	268.0 (44.67%)	0.32	False
Hypertension	844 (71.77%)	443.0 (73.83%)	0.39	False
Mortality	159 (13.52%)	206.0 (34.33%)	<0.001	True

This figure shows the t-test (for age) and chi-squared (for sex, diabetes, hypertension, and mortality) results between the COVID- cohort and the matched COVID+ cohort. We now see that there is no significant difference between any of the covariates (age, sex, diabetes, and hypertension)



In my Homework 3 feedback, it was suggested to me that I could do some correlation analysis with some of my data fields, but instead of just doing a regular correlation plot, I decided to make use of the KNeighborsRegressor. This is an especially useful tool because when Linear Regression is typically performed, it strictly uses a linear model to predict the association between two variables. However, with this form of regression, we are able to estimate the association between the variables based on the proximity of the neighboring data points, which is more effective since in the real-world, data is typically not close to normal. I chose to do correlations between age and heart rate and age and blood glucose, since part of my earlier analysis was to control for these variables since I knew they were likely to affect other data fields (such as mortality). My visualizations were two interactive plots showing these two correlations and also dividing the data points into males and females to add another layer of depth. Unfortunately, no statistically significant relationship was found between these variables, so I only included two of my attempts since they weren't significantly contributing to my research question.